

**Analyse en composantes principales et analyse
discriminante fonctionnelles appliquées à des
données de prises alimentaires animales**

par

Yann Décarie

**Mémoire présenté au département de mathématiques
en vue de l'obtention du grade de maître ès sciences (M.Sc.)**

**FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE**

Sherbrooke, Québec, Canada, 7 avril 2011



Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-79809-6
Our file Notre référence
ISBN: 978-0-494-79809-6

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

■ ■ ■
Canada

Le 10 mai 2011

*le jury a accepté le mémoire de Monsieur Yann Décarie
dans sa version finale.*

Membres du jury

Professeur Bernard Colin
Directeur de recherche
Département de mathématiques

Professeur Ernest Monga
Codirecteur de recherche
Département de mathématiques

Professeur Jean-Marc Belley
Membre
Département de mathématiques

Professeur Taoufik Bouezmarni
Président rapporteur
Département de mathématiques

Résumé

L'analyse de données fonctionnelles est une branche des statistiques modernes en pleine expansion. Cela est imputable aux avancées technologiques qui permettent et facilitent la collecte de large base de données ainsi que leurs représentations sous la forme de courbes ou de surfaces.

Ce mémoire se divise en trois parties. La première partie, consiste en la présentation des méthodes utilisées, soit l'analyse en composantes principales et l'analyse discriminante, dans le cadre de l'analyse des données classique. La seconde partie, a comme objectif de définir le cadre théorique permettant l'application de ces deux méthodes à des données fonctionnelles et également de présenter les résultats les plus pertinents à la compréhension des modèles. Enfin, la dernière partie porte sur les résultats de l'application de l'analyse en composantes principales et de l'analyse discriminante fonctionnelles aux données de prises alimentaires porcines. Le but de l'application de ces méthodes consiste à déterminer s'il est possible de différencier, seulement à l'aide des données disponibles, les porcs qui ont été malades des autres. Pour ce faire, une série de modèles, qui se distinguent par le choix du critère d'affectation à une classe donnée, ont été utilisés. On présentera les résultats des modèles jugés les plus pertinents permettant d'avoir une proportion de bien classés, supérieure à 85%.

Mots-clés : Analyse des données; analyse en composantes principales; analyse discriminante; analyse fonctionnelle; analyse de données fonctionnelles; analyse en composantes principales fonctionnelles; analyse discriminante fonctionnelle; données de prises alimentaires animales.

Remerciements

Je tiens à remercier l'Université de Sherbrooke et le département de mathématiques pour leur soutien autant financier que d'encadrement tout au long de la réalisation de mon mémoire. Je veux aussi remercier l'Institut des Sciences pour son soutien financier.

Je veux aussi remercier Taoufik Bouezmarni et Jean-Marc Belley pour avoir accepté de faire parti de mon jury et pour les améliorations que leurs commentaires ont apporté au mémoire.

Je remercie aussi mon codirecteur de recherche Ernest Monga pour son aide dans la réalisation de ce projet et Candido Pomar pour m'avoir laissé utiliser ses données.

Je remercie particulièrement mon directeur de recherche Bernard Colin sans qui ce mémoire n'aurait pas vu le jour. Aussi, pour sa confiance en mes capacités et cela dès notre première rencontre. Il m'a aussi fait comprendre que mathématique, statistique et plaisir peuvent très bien aller de pair.

Je veux également remercier mes parents Jean-Claude et Nicole qui m'ont toujours encouragé, supporté et aidé de façon inconditionnelle tout au long de mes études. Il m'est impossible d'exprimer dans un si court texte comment je vous suis redevable, mais sachez que j'en suis conscient.

Un merci bien spécial à Véronique pour sa compréhension et les sacrifices qu'elle a fait pour moi durant la réalisation de cette aventure.

Finalement, je remercie Gisèle pour ses corrections et Jacques pour m'avoir donné le goût aux études supérieures et à la recherche.

Table des matières

Résumé.....	4
Remerciements	5
Table des matières.....	6
Liste des tableaux.....	9
Liste des figures	10
I. Introduction	12
I.1 Analyse de données.....	12
I.2 Analyse de données fonctionnelles	13
I.3 Application.....	15
I.4 Plan du mémoire	16
Chapitre 1 : Présentation des méthodes dans le cadre de l'analyse de données multidimensionnelles classiques	17
1.1 Analyses multidimensionnelles	17
1.2 L'analyse en composantes principales (ACP)	18
1.2.1 Introduction.....	18
1.2.2 Données.....	19
1.2.3 Distance.....	20
1.2.4 Choix de l'origine.....	22
1.2.5 Inertie totale du nuage de point.....	24
1.2.6 Inertie du nuage des observations par rapport à un axe passant par G	25
1.2.7 Décomposition de l'inertie totale	26
1.2.8 Recherche du sous-espace vectoriel qui satisfait à la condition d'inertie minimale	29
1.2.9 Qualité de la représentation des observations	35
1.2.10 Représentation des variables	37
1.2.11 Qualité de la représentation des variables	40
1.2.12 Études des liaisons entre les variables	41
1.2.13 Observations et variables supplémentaires.....	43
1.2.14 Exemple	44
1.3 Analyse factorielle discriminante (AFD).....	54
1.3.1 Introduction	54
1.3.2 Données.....	56
1.3.3 Notations	58
1.3.4 Modèle	62
1.3.5 Métrique de Mahalanobis	64
1.3.6 Recherche des axes factoriels discriminants	64
1.3.7 Représentation des observations	66
1.3.8 Représentation des variables	67
1.3.9 Affectation d'une nouvelle observation à une classe k	68

1.3.10 Cas de 2 groupes	69
1.3.11 Évaluations des règles de décision	70
1.3.12 Exemple	71
Chapitre 2 : Outils pour l'analyse de données fonctionnelles	76
2.1 Dérivées et intégrales	79
2.2 Produit scalaire et espace L^p	80
2.2.1 L'espace L^p	80
2.2.2 Produit scalaire	82
2.3 Espace vectoriel	84
2.3.1 Espace vectoriel de dimension infinie	87
2.4 Statistiques descriptives	88
2.5 Espace de Hilbert	90
2.5.1 Définitions	90
2.5.2 Séparabilité et bases hilbertiennes	91
2.6 Espace de Sobolev	102
2.6.1 Dérivée faible	102
2.6.2 Espace de Sobolev	107
2.6.3 Théorème d'immersion	109
Chapitre 3 : Présentations des méthodes d'analyses de données dans le cadre fonctionnel	113
3.1 Introduction	113
3.2 Lien entre la théorie et la pratique	114
3.3 Analyse en composantes principales pour des données fonctionnelles	122
3.4 Analyse discriminante pour des données fonctionnelles	127
3.5 Résumé	130
Chapitre 4 : L'analyse des données	131
4.1 Introduction	131
4. 2 Méthodologie expérimentale	131
4.2.1 Animaux et régie	131
4.2.2 Aliments	132
4.3 Données	134
4.4 Transformation des variables	140
4.4.1 Les fonctions B-splines	142
4.4.2 Principales propriétés des fonctions B-splines	143
4.4.3 Approximation par des fonctions B-splines	145
4.4.4 Estimation	146
4.5 Lissage des données	149
4.6 Application de l'analyse en composantes principales fonctionnelles (ACPF) aux données de consommation cumulée porcine et à ses dérivées d'ordre un et deux	155
4.6.1 ACPF des courbes de consommation cumulée	155
4.6.2 ACPF de la première dérivée des courbes de consommation cumulée	162
4.6.2 ACPF de la deuxième dérivée des courbes de consommation cumulée	169

4.7 Analyse discriminante fonctionnelle de données de prises alimentaires	172
4.7.1 Détermination du critère de discrimination et de la nature des courbes à utiliser	173
4.7.2 Application de l'analyse discriminante fonctionnelle.....	180
Conclusion	185
Bibliographie	186

Liste des tableaux

Tableau 1: Présentation des données	45
Tableau 2: Centres de gravités et écarts-types	46
Tableau 3: Matrice de variance-covariance	47
Tableau 4: Valeurs propres de la matrice de variance-covariance	48
Tableau 5: Composantes principales pour chacune des années	50
Tableau 6: Qualité de la représentation	51
Tableau 7: Corrélations des nouvelles variables avec les anciennes	52
Tableau 8: Présentations des données	72
Tableau 9: Statistiques descriptives pour les trois groupes	73
Tableau 10: Valeurs propres des fonctions discriminantes	73
Tableau 11: Coefficients des fonctions discriminantes	74
Tableau 12: Tableau de classification des observations	75
Tableau 13 : Dates de début et de fin de la prise de mesure pour les 16 différents parcs	134
Tableau 14: Moyenne hebdomadaire de la consommation quotidienne (kg) pour les différents parcs	136
Tableau 15: Écart-type hebdomadaire de la consommation quotidienne (kg) pour les différents parcs	137
Tableau 16: Nombre de porcs malades et pourcentages selon différents critères	139

Liste des figures

Figure 1: Illustration de l'application du théorème de Pythagore.....	28
Figure 2 : Angle entre une observation et sa projection dans un plan	36
Figure 3: Cercle des corrélations pour les axes Δ_1 et Δ_2	38
Figure 4: Exemple d'un cercle des corrélations	42
Figure 5 : Projection du nuage dans le premier plan principal (plan 1-2)	49
Figure 6: Cercle des corrélations	53
Figure 7 : Illustration des données dans le cas de deux classes et de deux variables	55
Figure 8 : Illustration de l'axe discriminant dans le cas de deux classes et de deux variables.....	56
Figure 9: Représentation des observations dans le plan discriminant	74
Figure 10 : Exemples de données fonctionnelles	76
Figure 11: Schémas de dualité	120
Figure 12: Consommation alimentaire quotidienne (en kg) pour quatre porcs.....	138
Figure 13: consommation alimentaire cumulée (en kg) pour quatre porcs	138
Figure 14: Graphique tiré de l'exemple Ramsay sur des données de croissance humaine	140
Figure 15: Valeur de la validation croisée généralisée pour différence valeurs de lambda	151
Figure 16: Les courbes de consommation cumulée estimées pour les 177 porcs	152
Figure 17: Erreur quadratique moyenne pour chacun des 177 porcs	152
Figure 18: Approximation et résidus pour les deux porcs ayant un EQM extrême	153
Figure 19: Agrandissement du graphique de la fonction d'approximation pour un porc.....	154
Figure 20: Les 10 premières valeurs propres de l'analyse en composantes principales fonctionnelles de la consommation cumulée	156
Figure 21: Première et de la deuxième composantes principales.....	156
Figure 22: Effet de la première composante principale sur la moyenne de la consommation cumulée	157
Figure 23 : Effet de la deuxième composante principale sur la moyenne de la consommation cumulée	158
Figure 24: Projection des observations dans le plan constitué par le premier et le deuxième axe principal	159
Figure 25 : Projections des Porcs près de l'origine du plan constitué par le premier et le deuxième axe principal	160
Figure 26: Consommation cumulée de deux porcs bien discriminés et de deux autres ayant des coordonnées proches selon le premier axe principal.....	161

Figure 27: Quatre porcs ayant des coordonnées extrêmes par rapport à la deuxième composante principale.....	162
Figure 28: Dérivée première de la consommation alimentaire cumulée pour tous les porcs	163
Figure 29 : Les 10 premières valeurs propres de l'analyse en composantes principales fonctionnelles de la première dérivée des courbes de consommation cumulée	164
Figure 30: Première et deuxième composantes principales pour L'ACPF de la première dérivée des courbes de consommation cumulée.....	165
Figure 31: Effet de la première composante principale sur la moyenne	165
Figure 32 : Effet de la deuxième composante principale sur la moyenne.....	166
Figure 33 : Projection des observations dans le plan constitué par le premier et le deuxième axe principal	167
Figure 34 : Projections des observations près de l'origine du plan constitué par le premier et le deuxième axe principal	168
Figure 35: Dérivée seconde de la consommation alimentaire cumulée pour tous les porcs	169
Figure 36 : Les 20 premières valeurs propres de l'analyse en composantes principales de la seconde dérivée des courbes de consommation cumulée....	170
Figure 37 : Pourcentage de variabilité expliquée par les 20 premières valeurs propres.....	171
Figure 38 : Valeurs de la première composante principale pour L'ACPF de la deuxième dérivée des courbes de consommation cumulée	172
Figure 39 : Moyenne des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les différents critères.....	174
Figure 40 : La somme des écarts entre les courbes moyenne de consommation cumulée des porcs en santé et des porcs malade selon les sept critères	175
Figure 41 : Moyenne de la première dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les différents 4+ à 7+.....	176
Figure 42 : Moyenne de la première dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 1+ à 3+.....	177
Figure 43 : La somme des écarts entre la première dérivée des courbes moyennes de consommation cumulée des porcs en santé et des porcs malades selon les sept critères	177
Figure 44 : Moyenne de la deuxième dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 4+ à 7+.....	179
Figure 45 : Moyenne de la deuxième dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 1+ à 3+.....	179
Figure 46: Consommation cumulée des porcs malades et la fonction discriminante sous le critère 6+	180
Figure 47 : Classification <i>a posteriori</i> des observations selon le critère 6+.....	181

Figure 48 : Classification <i>a posteriori</i> des dérivées premières selon le critère 6+	182
Figure 49 : Classification <i>a posteriori</i> des dérivées premières selon le critère 3+	184

I. Introduction

La statistique est une branche trop souvent méconnue des mathématiques. Elle nous permet, grâce à l'analyse des données, de traiter de façon objective et de comprendre les liens qui unissent les différentes variables observées. De nos jours, il est de plus en plus facile d'avoir accès à une masse considérable d'information sur un sujet donné. Les avancées continues des nouvelles technologies permettent de mesurer une multitude de variables plus rapidement, plus fréquemment et à moindres coûts. Il est aussi plus facile de les conserver pour un traitement éventuel. Pour ces raisons, les méthodes d'analyses doivent aussi évoluer et s'adapter à ces nouvelles réalités. L'analyse de données fonctionnelles est une méthode récente qui permet de faire face à ces nouveaux défis en analysant non plus des points dans \mathcal{R}^n , mais des courbes ou des fonctions dans un espace de dimension infinie. C'est sous ce thème que s'inscrivent les travaux effectués dans le cadre de ce mémoire.

I.1 Analyse de données

La statistique, comme elle est définie aujourd'hui, est une discipline très jeune dans l'histoire des sciences. Bien qu'il y ait traces de recensement (décompte) de la population, des terres agricoles ou du bétail depuis la création des premières structures sociales, ce n'est que depuis le XIX^e siècle que la discipline a vraiment pris son envol. L'analyse des données est une branche encore plus récente. Elle a vu le jour au début du XX^e siècle pour permettre la compréhension de données multidimensionnelles grandement facilitées par le développement des dernières années en termes d'outils techniques permettant d'effectuer, à l'intérieur de délais et d'effort raisonnable, des calculs matriciels à

ce jour inimaginables. L'analyse de données a pour but de synthétiser l'information, au sens de la « théorie de l'information », pour en faciliter la compréhension et l'interprétation de divers phénomènes mesurables. Prenons comme exemple des données dans \mathfrak{R}^n que l'on projette dans un espace à deux dimensions sur les axes où l'inertie des données est la plus grande possible. Cela permet de distinguer rapidement quelles variables caractérisent le mieux l'ensemble des données. Ceci n'est qu'une illustration élémentaire des multiples possibilités qu'offre ce domaine de la statistique et elle sera expliquée en détail dans le premier chapitre.

1.2 Analyse de données fonctionnelles

L'analyse de données fonctionnelles prend racine dans l'analyse fonctionnelle classique, qui est l'étude de l'espace des fonctions et de l'analyse de données classique. C'est un domaine d'étude encore jeune dû aux outils techniques, tels ordinateurs et logiciels, nécessaires à sa réalisation et aussi à la quantité de données maintenant disponibles dans des temps de plus en plus courts. Cela nous permet d'avoir un regard plus juste sur des phénomènes qui sont de nature continue, comme la prise de poids d'un individu ou la température d'un endroit donné, et qui étaient discrétisés par manque d'information ou tout simplement pour en faciliter l'analyse. Ce champ des statistiques est porté à prendre de l'expansion, car la disponibilité de données en grande quantité est une réalité qui ne va que s'amplifier au fil du temps. Bien que l'analyse de données fonctionnelles couvre plusieurs méthodes, les travaux effectués dans le cadre de ce mémoire vont être axés sur l'analyse en composantes principales de données fonctionnelles, l'analyse discriminante de données fonctionnelles et appliquées à des données de prise alimentaire porcine.

Dans la littérature moderne, on retrouve plusieurs articles et auteurs importants dans le développement de l'analyse de données fonctionnelles. On présente ici une brève liste de ceux qui ont été prépondérants dans la réalisation de ce

mémoire. On commence par citer un texte de Ramsay [Ramsay *et al.*, 1991] qui, en plus de présenter clairement les grandes lignes de l'analyse de données fonctionnelle, démontre pourquoi les L-splines cadrent dans la théorie des modèles linéaires généralisés et dans celle de l'analyse fonctionnelle en composantes principales. L'article de Silverman [Silverman 1996] est important car il présente une façon de prendre en compte certaines contraintes des fonctions que l'on désire traiter dans le cadre de l'analyse en composantes principales. Il y a aussi, l'article de Lee [Lee *et al.* 2002] qui se concentre sur une méthode en deux étapes pour le calcul des fonctions de covariance, des valeurs propres et des fonctions propres dans le cadre de données fonctionnelles. Un autre incontournable est le livre de Ramsay [Ramsay *et al.* 2002]. Ce manuel porte sur l'application de différentes méthodes utilisées en analyse de données fonctionnelles. L'une des études de cas présentée au chapitre six du livre est particulièrement intéressante pour nous, car elle porte sur des données concernant la croissance humaine ce qui est similaire au domaine d'application qui sera considéré dans ce mémoire. Dans le même ordre d'idée, le livre de Dado-Niaug [Dabo-Niaug 2008] contient plusieurs applications de l'analyse de données fonctionnelles.

Sur le plan plus technique, il y a le texte de Clarkson [Clarkson *et al.* 2005], qui présente les fonctions spécifiques au logiciel S-PLUS et donc à son pendant gratuit R, de l'analyse de données fonctionnelles. Finalement, il y a la seconde édition du livre de Ramsay [Ramsay *et al.* 2005] qui s'efforce de couvrir le cadre théorique de l'analyse de données fonctionnelles tout en gardant une approche pédagogique intéressante.

Il faut aussi citer le livre de Ferraty [Ferraty *et al.* 2006] qui adopte un point de vue non-paramétrique. On sait que les méthodes non-paramétriques ont fait leurs preuves en les espaces de petites dimensions, notamment dans les problèmes de prévision. En grandes dimensions et *a fortiori* en dimension infinie, leur emploi est plus délicat à cause de la raréfaction locale des observations. Les

auteurs s'attaquent à ce problème crucial et le résolvent d'une façon convaincante en introduisant des hypothèses adéquates et réalistes.

Cette brève recension des écrits est loin d'être exhaustive, mais elle a le mérite de faire une entrée en matière claire et précise de ce qu'est l'analyse de données fonctionnelles. De plus, on trouve, à la section bibliographie, une liste complète des textes consultés pour la rédaction de ce mémoire.

I.3 Application

Ce mémoire n'est pas seulement théorique, on y retrouve une application sur des données de prise alimentaire porcine qui sera présentée au chapitre 4. Les domaines de l'agriculture et de l'agroalimentaire ont beaucoup changé. Pour atteindre et soutenir un niveau de rentabilité, les producteurs doivent atteindre de nouveaux seuils d'efficacités. Pour les aider dans cette démarche, les nouvelles technologies sont des outils quasi indispensables et ce sont les données recueillies par ces nouveaux systèmes que nous nous proposons d'analyser.

Les données proviennent d'un élevage de porcs de la ferme expérimentale de Lennoxville [Pomar 2007]. Le protocole et les instruments mis en place permettent de recueillir la quantité de nourriture que chacun des porcs de l'élevage ingurgite dans un intervalle de temps donné. Une fois les données transformées en fonctions ou courbes, on tentera de déterminer si les variations observées en termes de prise alimentaire quotidienne, peuvent servir à détecter une maladie chez le porc en question. Si l'analyse fonctionnelle se révèle efficace, elle permettra d'isoler l'individu malade afin de diminuer le risque de contagion. Ces travaux sont une continuité de ceux effectués par K. Chenguiti dans le cadre de son mémoire de maîtrise, [Chenguiti 2007].

I.4 Plan du mémoire

Ce mémoire est divisé en quatre chapitres. Le premier, porte sur la présentation de l'analyse en composantes principales et l'analyse discriminante dans le cadre de données classiques. Le second a pour but de mettre en place les outils théoriques essentiels à l'analyse de données fonctionnelles. Le troisième est le développement des deux méthodes exposées au premier chapitre, mais dans le cadre de données fonctionnelles. Le quatrième est l'application des méthodes dans le cas de données de prise alimentaire porcine.

Chapitre 1 : Présentation des méthodes dans le cadre de l'analyse de données multidimensionnelles classiques

1.1 Analyses multidimensionnelles

L'analyse de données multidimensionnelles a pour but de représenter de façon optimale, façon qui sera définie ultérieurement, des objets caractérisés par des variables descriptives et ce à l'aide de différents critères objectifs. Cette analyse englobe plusieurs méthodes entièrement basées sur la représentation des observations par un nuage de points dans l'espace vectoriel des observations (ou son dual). Elle veut aussi résumer l'information, au sens de la « théorie de l'information » [Kullback 1959], contenue dans le nuage de points en utilisant un nombre restreint des variables disponibles ou plus généralement des combinaisons linéaires de celles-ci. Un des avantages marqués des méthodes d'analyse qui sont présentées dans ce chapitre est de ne faire aucune hypothèse sur les lois de densités de probabilité des objets analysés contrairement aux méthodes d'analyse statistique multidimensionnelle. Elles sont dites non-paramétriques. Ce type d'hypothèse, comme la normalité ou la multi-normalité des données, est contraignante et n'est pas toujours valable ou facilement vérifiable dans un espace de grande dimension. Un autre avantage réside dans la possibilité de projeter le nuage de points dans un espace à deux ou trois dimensions, ce qui facilite la compréhension intuitive des données.

On peut distinguer deux grandes classes de problèmes traités en analyse de données. Premièrement, il y a les approches de nature descriptive qui ont pour but de faire ressortir les liens entre les différents objets étudiés. Pour résoudre ce type de problèmes, on utilise différentes méthodes telles l'analyse en

composantes principales ou l'analyse factorielle des correspondances. Deuxièmement, il y a les approches de nature décisionnelle qui elles consistent, comme leur nom l'indique, à prendre, entre autres, une décision sur la classe d'appartenance observable d'un objet analysé. Pour ce faire, on peut utiliser l'analyse discriminante. Bien que l'analyse des données englobe une multitude de méthodes, on présentera uniquement l'analyse en composantes principales et l'analyse discriminante car se sont les seules qui seront utilisées dans ce mémoire.

1.2 L'analyse en composantes principales (ACP)

1.2.1 Introduction

Bien que les pères fondateurs des méthodes factorielles soient le britannique Charles E. Spearman [Spearman 1904] et l'Américain Louis L. Thurstone [Thurstone 1941] de par leurs travaux dans le domaine de la psychométrie, on doit l'analyse en composantes principales à Karl Pearson qui en publia les grandes lignes [Pearson 1901] ainsi qu'à Harold Hotelling qui formalisa le tout en 1933 [Hotelling 1933]. Bien que ces publications datent du début de XXe siècle, il a fallu attendre l'arrivée des outils informatiques plus performants, vers 1960, pour que cette technique, surtout développée en France particulièrement par Jean-Paul Benzécri [Benzécri 1969], prenne réellement son envol.

Selon Crucianu [Crucianu *et al.* 2004] l'ACP se définit comme suit : « C'est une méthode d'analyse multidimensionnelle exploratoire : à partir d'une population d'observations caractérisées par un ensemble de variables quantitatives initiales, on cherche à structurer et à résumer les variables et les observations afin de mieux appréhender l'information apportée par les variables. L'ACP fournit un moyen de condenser la représentation des données en conservant au mieux leur organisation globale. ». Cette méthode résume l'information contenue dans l'ensemble des variables initiales en utilisant les composantes principales

obtenues par des combinaisons linéaires des variables initiales. Pour faciliter la compréhension des données, on peut projeter les observations dans un espace de plus petite dimension, \mathcal{R}^2 ou \mathcal{R}^3 , où les axes sont, par exemple, les composantes principales. Il est aussi important de noter que l'ACP peut servir d'étape préliminaire à une analyse discriminante en éliminant les variables redondantes et en identifiant les axes selon lesquels l'inertie des données est la plus élevée. Par opposition aux analyses confirmatoires (par exemple l'analyse de variance), qui visent à tester des hypothèses fondées sur un raisonnement théorique, l'ACP permet de découvrir les relations entre variables sans avoir recours à des hypothèses préalables de nature probabiliste.

L'ACP n'est pas invariante aux changements d'échelles des variables utilisées, elle est donc sensible aux unités de mesures utilisées. Il est important, si les colonnes d'une matrice contiennent des données mettant en jeu des variables hétérogènes telles que celles définie à la sous-section suivante (par exemple : des mètres et des kilogrammes), de les normaliser afin d'avoir un résultat indépendant des unités de mesures utilisées. Si en revanche les colonnes sont comparables, on peut préférer de ne pas faire de normalisation : on considère ainsi que l'information de niveau relatif entre les différentes variables est importante; les variables de faibles amplitudes seront pénalisées car elles interviendront moins dans les premières composantes principales.

Enfin, comme pour toute méthode descriptive, réaliser une ACP n'est pas une fin en soi. L'ACP servira à mieux connaître les données sur lesquelles on travaille, à détecter des valeurs suspectes ou aberrantes. De plus, elle aidera à formuler des hypothèses à l'aide de modèles et d'études statistiques inférentielles. On pourra aussi, *a posteriori*, se servir des représentations fournies par l'ACP pour illustrer certains résultats dans un but pédagogique.

1.2.2 Données

L'ACP s'applique à des données quantitatives contenues dans une matrice rectangulaire X à n lignes et p colonnes. Chaque ligne $i : 1, 2, \dots, n$ représente une observation x_i et chaque colonne $j : 1, 2, \dots, p$ représente une variable v_j .

$$X = \{x_{ij}\}_{i:1,2,\dots,n; j:1,2,\dots,p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \text{ où } x_{ij} \in \mathbb{R}$$

On peut aussi représenter la matrice X en termes des observations x_i ou des variables v_j qui la composent :

$$x_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T \quad x_i \in \mathbb{R}^p$$

$$v_j = [x_{1j}, x_{2j}, \dots, x_{nj}]^T \quad v_j \in \mathbb{R}^n$$

1.2.3 Distance

Pour pouvoir effectuer une représentation géométrique, il faut choisir une distance entre deux points de l'espace. Une distance qui est fréquemment utilisée lorsque l'on effectue une ACP dans l'espace où sont représentées les observations, est la distance euclidienne classique notée d . La distance entre deux observations x_i et x_j est alors égale à :

$$d(x_i, x_j) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$

Par contre, il est important de noter que cette même distance peut être calculée à partir de n'importe laquelle des métriques associées à une matrice A symétrique et définie positive. Dans ce cas, le carré de la distance entre deux observations x_i et x_j est égale à :

$$d^2(x_i, x_j) = (x_i - x_j)^T A (x_i - x_j)$$

Sans perte de généralité et surtout pour faciliter la compréhension du texte, la suite de la description des méthodes sera effectuée en prenant la distance euclidienne classique comme référence, à moins de mention du contraire.

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. À cette distance, on associe un produit scalaire entre deux vecteurs :

$$\langle x_i, x_j \rangle = \sum_{k=1}^p x_{ik} x_{jk} = x_i^T x_j$$

Le choix de l'origine O sera discuté au point suivant. De ce produit scalaire, on peut aussi en déduire la norme suivante pour un vecteur:

$$\sum_{k=1}^p x_{ik}^2 = x_i^T x_i = \|x_i\|^2$$

On peut alors définir l'angle α entre deux vecteurs par son cosinus :

$$\cos(\alpha) = \frac{\langle x_i, x_j \rangle}{\|x_i\| \|x_j\|} = \frac{\sum_{k=1}^p x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^p x_{ik}^2} \sqrt{\sum_{k=1}^p x_{jk}^2}} = \frac{x_i^T x_j}{\sqrt{(x_i^T x_i)(x_j^T x_j)}}$$

1.2.4 Choix de l'origine

Le point 0, correspondant au vecteur dont toutes les coordonnées sont nulles n'est pas forcément une origine satisfaisante, car si les coordonnées des points du nuage N des observations sont grandes, le nuage est éloigné de cette origine. Il apparaît plus judicieux de choisir une origine liée au nuage lui-même : le centre de gravité du nuage par exemple. Pour définir ce centre de gravité, il faut choisir un système de pondération pour les observations. Les poids p_i , $\forall i = 1, 2, \dots, n$ des observations x_i sont généralement choisis tel que :

$$\sum_{i=1}^n p_i = 1 \quad \text{avec } p_i > 0 \quad \forall i = 1, 2, \dots, n$$

Ainsi N peut s'écrire de la façon suivante :

$$N = \left\{ (x_i, p_i) : x_i \in \mathbb{R}^p; p_i > 0 \forall i = 1, 2, \dots, n; \sum_{i=1}^n p_i = 1 \right\}$$

Le centre de gravité G est défini comme le vecteur de \mathbb{R}^p tel que :

$$G = \sum_{i=1}^n p_i x_i$$

et donc

$$\sum_{i=1}^n p_i (x_i - G) = 0$$

Par ailleurs, il est fréquent de noter G par \bar{X} où la $j^{\text{ème}}$ composante est donnée par :

$$\bar{x}_j = \sum_{i=1}^n p_i x_{ij}$$

Si l'on choisit de donner le même poids $1/n$ à toutes les observations, le centre de gravité G du nuage des observations est alors le vecteur dont les coordonnées sont les valeurs moyennes des variables que l'on écrit comme suit :

$$G = \begin{pmatrix} \sum_{i=1}^n p_i x_{i1} \\ \vdots \\ \sum_{i=1}^n p_i x_{ij} \\ \vdots \\ \sum_{i=1}^n p_i x_{ip} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

Prendre G comme origine revient alors à travailler sur le tableau des données centrées X_c :

$$X_c = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

1.2.5 Inertie totale du nuage de point

On note I_G l'inertie du nuage des observations par rapport au centre de gravité G que l'on définit comme suit:

$$I_G = \sum_{i=1}^n p_i d^2(G, x_i) = \sum_{i=1}^n p_i \sum_{j=1}^p (x_{ij} - \bar{x}_j)^2 = \sum_{i=1}^n p_i x_{ci}^T x_{ci}$$

où

$$x_{ci} = \begin{pmatrix} x_{i1} - \bar{x}_1 \\ \vdots \\ x_{ij} - \bar{x}_j \\ \vdots \\ x_{ip} - \bar{x}_p \end{pmatrix}$$

L'inertie totale est intéressante car c'est une mesure de la dispersion du nuage des observations par rapport à son centre de gravité. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré autour de son centre de gravité. Dans le cas limite où l'inertie est nulle, tous les points sont confondus.

On peut aussi écrire I_G de façon différente en permutant l'ordre des deux sommations, ce qui permet de faire le lien entre l'inertie du nuage et la variance empirique des variables v_j qui compose ce même nuage :

$$I_G = \sum_{j=1}^p \left[\sum_{i=1}^n p_i (x_{ij} - \bar{x}_j)^2 \right] = \sum_{j=1}^p \text{Var}(v_j)$$

Avec cette écriture, on remarque de l'inertie totale est la trace de la matrice de variance-covariance Σ des p variables définie par :

$$\Sigma = \begin{pmatrix} \text{Var}(v_1) & \text{Cov}(v_1, v_2) & \cdots & \text{Cov}(v_1, v_j) & \cdots & \text{Cov}(v_1, v_p) \\ \text{Cov}(v_2, v_1) & \text{Var}(v_2) & \cdots & \text{Cov}(v_2, v_j) & \cdots & \text{Cov}(v_2, v_p) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \text{Cov}(v_j, v_1) & \text{Cov}(v_j, v_2) & \cdots & \text{Var}(v_j) & \cdots & \text{Cov}(v_j, v_p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(v_p, v_1) & \text{Cov}(v_p, v_2) & \cdots & \text{Cov}(v_p, v_j) & \cdots & \text{Var}(v_p) \end{pmatrix}$$

où

$$\text{Cov}(v_i, v_j) = \sum_{k=1}^n p_k [(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)]$$

1.2.6 Inertie du nuage des observations par rapport à un axe passant par G

L'inertie du nuage des observations par rapport à un axe Δ passant par G , notée I_Δ est définie de la façon suivante :

$$I_{\Delta} = \sum_{i=1}^n p_i d^2(h_{\Delta_i}, x_i)$$

où h_{Δ_i} représente la projection orthogonale de x_i sur l'axe Δ . Cette inertie mesure la proximité du nuage de points des observations autour l'axe Δ .

Sans perte de généralité, on considère dans ce mémoire que le centre de gravité G est toujours confondu avec l'origine, car il est toujours possible, par une simple translation, de se ramener à ce cas.

1.2.7 Décomposition de l'inertie totale

Si on note par V^{\perp} le sous-espace supplémentaire orthogonal de V , un sous-espace de \mathbb{R}^p , et $h_{V^{\perp}_i}$ la projection orthogonale de x_i sur V^{\perp} , en appliquant le théorème de Pythagore, on peut écrire :

$$d^2(h_{V_i}, x_i) + d^2(h_{V^{\perp}_i}, x_i) = d^2(G, x_i) = d^2(G, h_{V_i}) + d^2(G, h_{V^{\perp}_i})$$

$$d^2(h_{V_i}, x_i) + d^2(h_{V^{\perp}_i}, x_i) = \|x_i\|^2 = \|h_{V_i}\|^2 + \|h_{V^{\perp}_i}\|^2$$

On en déduit que l'inertie totale peut s'écrire comme suit :

$$I_G = I_V + I_{V^{\perp}}$$

Dans le cas particulier où le sous-espace V est de dimension 1, c'est-à-dire dans le cas où V est un axe, $I_{V^{\perp}}$ est une mesure de l'allongement du nuage selon cet axe. On emploie parfois pour définir $I_{V^{\perp}}$ l'expression « inertie

expliquée par l'axe V ». En projetant le nuage des observations sur un sous-espace V , on perd l'inertie mesurée par I_V , on ne conserve que celle mesurée par I_{V^\perp} .

De plus, si on décompose l'espace \mathfrak{R}^p comme la somme directe de sous-espaces de dimension 1 orthogonaux deux à deux :

$$\Delta_1 \oplus \Delta_2 \oplus \dots \oplus \Delta_p$$

on peut écrire I_G comme étant :

$$I_G = \sum_{i=1}^n p_i d^2(G, x_i) = I_{\Delta_1} + I_{\Delta_2} + \dots + I_{\Delta_p}$$

On peut aussi appliquer le théorème de Pythagore, ce qui donne :

$$d^2(G, x_i) = d^2(G, h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i}) + d^2(x_i, h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i})$$

où $h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i}$ est la projection orthogonale de x_i sur le sous-espace Δ_1^\perp , le supplémentaire orthogonal de Δ_1 dans \mathfrak{R}^p . Et comme par définition, tous les axes sont orthogonaux, cela revient à dire que :

$$d^2(x_i, h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i}) = d^2(G, h_{\Delta_1 i})$$

La figure 1 fourni une illustration simple de l'application du théorème de Pythagore. Dans cette illustration, on remarque que :

$$\|x_i - G\|^2 = \|x_i - h_{\Delta_i}\|^2 + \|h_{\Delta_i} - G\|^2$$

Il est donc clair que trouver les valeurs de h_{Δ_i} les plus proches de celles de x_i dans un nouvel espace, revient à maximiser la dispersion (ou inertie totale) des h_{Δ_i} .

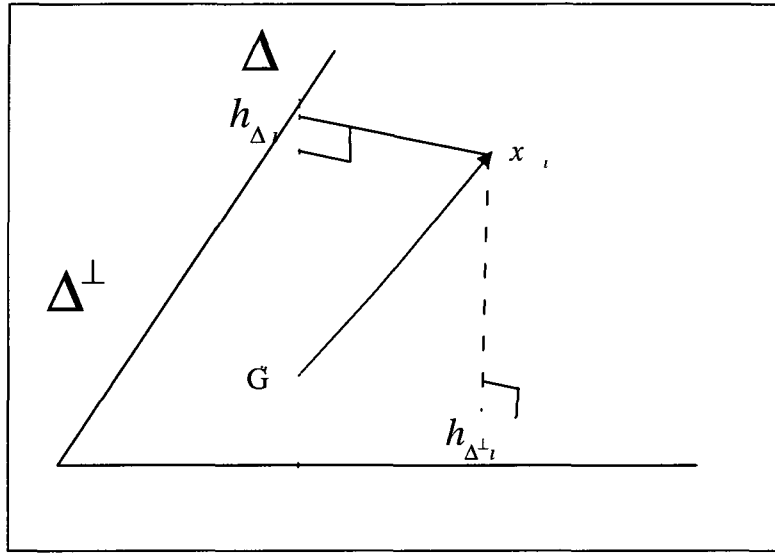


Figure 1: Illustration de l'application du théorème de Pythagore

De la même façon, on peut réécrire $d^2(G, h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i})$ comme suit :

$$d^2(G, h_{(\Delta_2 \oplus \dots \oplus \Delta_p)_i}) = d^2(G, h_{(\Delta_3 \oplus \dots \oplus \Delta_p)_i}) + d^2(G, h_{\Delta_2 i})$$

Après $p - 1$ décompositions identiques, on arrivera au résultat suivant :

$$d^2(G, x_i) = d^2(G, h_{\Delta_1 i}) + d^2(G, h_{\Delta_2 i}) + \dots + d^2(G, h_{\Delta_p i})$$

En utilisant cette expression du carré de la distance de x_i au centre de gravité G , on retrouve la décomposition de l'inertie totale comme somme des inerties expliquées par les p axes orthogonaux.

$$I_G = I_{\Delta_1} + I_{\Delta_2} + \dots + I_{\Delta_p}$$

1.2.8 Recherche du sous-espace vectoriel qui satisfait à la condition d'inertie minimale

L'objectif de l'ACP consiste à déterminer un sous-espace linéaire $V \in \mathbb{R}^q$, avec $q < p$, de \mathbb{R}^p , l'espace des observations, tel que l'inertie des observations projetées sur V soit maximale ou encore qui minimise l'erreur faite lors de l'approximation du nuage de points X par sa projection sur V . On cherche donc V tel que I_V soit minimum ou encore tel que I_{V^\perp} soit maximum. À des fins de simplification de la notation, les démarches de cette section seront présentées pour la matrice des données centrées X .

On doit donc trouver un sous-espace qui satisfasse à ces conditions. Nous définissons un tel sous-espace de dimension q par l'ensemble des q vecteurs unitaires et orthogonaux notés $\mu_\alpha, \alpha: 1, 2, \dots, q$. Ces q vecteurs forment une base orthonormée du sous-espace. De plus, on montre dans [Crucianu *et al.* 2004] que ces q vecteurs sont les vecteurs propres de $X^T X$ (où X^T représente la matrice X transposée) associés aux q plus grandes valeurs propres de cette même matrice. Les vecteurs μ_α satisferont alors l'équation suivante :

$$X^T X \mu_\alpha = \lambda_\alpha \mu_\alpha$$

Aussi, (voir [Rotella *et al.* 1995]) on peut affirmer, en vertu des propriétés de la matrice $X^T X$, que les vecteurs propres μ_α associés aux valeurs propres non nulles différentes λ_α sont orthogonaux. L'indice α d'une valeur propre représente sa position dans la suite ordonnée $\lambda_1 > \lambda_2 > \dots > \lambda_i > \dots > \lambda_q > 0$.

Comme ce problème d'optimisation est à la base de l'analyse en composantes principales, il a été jugé bon de poser et de résoudre le problème dans un cadre général, soit sous une métrique associée à la matrice A symétrique définie-positive. Il faut donc trouver le vecteur u tel que I_{V^\perp} soit maximale :

$$\max_{u \in \mathbb{R}^p} u^T A V_N A u$$

où $V_N = X^T D X$ est dite la matrice d'inertie du nuage de points N et D la matrice diagonale des poids p_i .

Sous la contrainte suivante qui force le vecteur à être de longueur unitaire :

$$\|u\|^2 = u^T A u = 1$$

On peut trouver la solution de ce problème d'optimisation en appliquant la méthode des multiplicateurs de Lagrange :

$$\varphi(u, \lambda) = u^T A V_N A u - \lambda (u^T A u - 1)$$

$$\frac{d}{d\lambda} [u^T A V_N A u - \lambda (u^T A u - 1)] = u^T A u - 1 = 0$$

donc, comme la dérivée par rapport à λ est égale à 0, car $u^T A u = 1$ on doit résoudre seulement l'expression suivante :

$$\frac{d}{du} [u^T A V_N A u - \lambda (u^T A u - 1)] = 0$$

c'est-à-dire :

$$A V_N A u - \lambda A u = 0$$

Comme A est symétrique définie-positive, A est inversible :

$$V_N A u - \lambda u = 0$$

Donc, u est vecteur propre de la matrice $V_N A$ associé à la valeur propre λ .

D'autre part pour le vecteur propre u correspondant à la valeur propre λ , on a :

$$I_{V^\perp} = u^T A V_N A u \geq 0$$

$$I_{V^\perp} = u^T A \lambda u = \lambda$$

Étant donné que l'on cherche à maximiser I_{V^\perp} , u est donc le vecteur propre de $V_N A$ correspondant à la plus grande valeur propre λ_1 . On note u_1 ce vecteur propre, vecteur unitaire de l'axe Δ_{u_1} .

Si l'on cherche ensuite un deuxième axe Δ_{u_2} orthogonal, ou A -orthogonal dans le cas général, à Δ_{u_1} et tel que $I_{\Delta_{u_2}}$ soit minimum, on est amené à résoudre le système d'équations suivant :

$$\max_{u \in \mathfrak{R}^p} u^T A V_N A u$$

$$\|u\|^2 = u^T Au = 1$$

et la contrainte d'orthogonalité :

$$\langle u, u_1 \rangle = u^T Au_1 = 0$$

En utilisant encore la méthode des multiplicateurs de Lagrange, on obtient u solution de :

$$\varphi(u, \lambda, v) = u^T AV_N Au - \lambda (u^T Au - 1) - v u^T Au_1$$

$$\frac{d}{d\lambda} [u^T AV_N Au - \lambda (u^T Au - 1) - v u^T Au_1] = u^T Au - 1 = 0$$

$$\frac{d}{dv} [u^T AV_N Au - \lambda (u^T Au - 1) - v u^T Au_1] = u^T Au_1 = 0$$

donc, comme la dérivée par rapport à λ est égale à 0, car $u^T Au = 1$ et que la dérivée par rapport à v est aussi égale à zéro, car $u^T Au_1 = 0$, on doit résoudre seulement l'expression suivante :

$$\frac{d}{du} [u^T AV_N Au - \lambda (u^T Au - 1) - v u^T Au_1] = 0$$

en dérivant l'équation précédente, on obtient :

$$2AV_N Au - 2\lambda Au - vAu_1 = 0$$

en prémultipliant à gauche par u_1^T on obtient :

$$2u_1^T AV_N Au - 2\lambda u_1^T Au - vu_1^T Au_1 = 0$$

comme $V_N Au = \lambda u$, ce qui implique que $u_1^T V_N A = u_1^T \lambda$ et donc :

$$\begin{aligned} 2\lambda u_1^T Au - 2\lambda u_1^T Au - vu_1^T Au_1 &= 0 \\ vu_1^T Au_1 &= 0 \end{aligned}$$

on en déduit que $v = 0$ car on sait que $u_1^T Au_1 = 1$. Ce qui nous ramène au même problème que pour le premier axe :

$$\begin{aligned} AV_N Au - \lambda Au &= 0 \\ V_N Au &= \lambda u \end{aligned}$$

Ce qui veut dire que u_2 , vecteur unitaire de Δ_{u_2} , est donc le vecteur propre de $V_N A$, orthogonal ou A -orthogonal à u_1 , correspondant à la deuxième plus grande valeur propre λ_2 .

De manière générale, on peut montrer que l'hyperplan $V \in \mathfrak{R}^q \subseteq \mathfrak{R}^p$ tel que I_V soit minimum, possède une base orthonormée ou A -orthonormée constituée des q vecteurs propres associés au q plus grandes valeurs propres de $V_N A$.

Maintenant que le problème d'optimisation a été formalisé et résolu dans un cadre général, la suite est présentée dans le cas de la métrique euclidienne usuelle. Le prolongement du vecteur unitaire μ_α est utilisé comme droite de projection pour les observations x_i et est appelée le $\alpha^{ième}$ axe principal d'inertie ou l'axe factoriel d'ordre α . La projection orthogonale de l'observation x_i sur l'axe principal d'ordre α est donnée par :

$$\psi_{\alpha i} = x_i^T \mu_{\alpha}$$

L'inertie des projections sur l'axe principal α est donnée par :

$$\mu_{\alpha}^T X^T X \mu_{\alpha} = \sum_{i=1}^n \psi_{\alpha i}^2 = \lambda_{\alpha} = I_{\Delta_{\alpha}}^{\perp}$$

L'inertie totale de la projection des observations sur le sous-espace des q premiers axes factoriels quant à elle, est donnée par la somme de l'inertie des projections sur chaque axe qui compose ce sous-espace, soit les q plus grandes valeurs propres de $X^T X$:

$$I_V = \sum_{i=1}^q \lambda_i$$

En combinant les deux derniers résultats, on en déduit une expression pour la contribution relative de l'axe principal α par rapport à l'inertie totale du nuage de point ou le pourcentage d'inertie expliquée par l'axe principal α :

$$\frac{I_{\alpha}}{I_V} = \frac{\lambda_{\alpha}}{\sum_{i=1}^p \lambda_i}$$

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliquée par le plan engendré par les deux premiers axes principaux, $\Delta_{1 \oplus 2}$ est égal à :

$$\frac{I_{\Delta_{1 \oplus 2}}}{I_V} = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des observations expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

On se contente souvent de faire des représentations du nuage de points dans un sous-espace engendré par les d premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse à un sous-espace de dimension $d \leq q \leq p$.

1.2.9 Qualité de la représentation des observations

Lorsque les projections des observations sont éloignées sur un axe (ou dans un plan), on peut assurer que les points représentant ces observations sont éloignés dans l'espace \mathfrak{R}^p . En revanche, deux observations dont les projections sont proches sur un axe (ou dans un plan ou dans un sous-espace) peuvent ne pas être proches dans l'espace \mathfrak{R}^p . Pour interpréter correctement la proximité des projections de deux observations sur un plan, il faut donc s'assurer que ces observations soient bien représentées dans le plan.

Pour que l'observation x_i soit bien représenté sur un axe (ou sur un plan, ou un sous-espace), il faut que l'angle entre le vecteur x_i et l'axe (ou le plan, ou le sous-espace) soit petit. On calcule donc le cosinus de cet angle, ou plutôt le carré de ce cosinus. En effet, en utilisant de nouveau le théorème de Pythagore, on remarque que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. La figure 2 illustre l'équation suivante :

$$\cos^2(\theta) = \cos^2(\theta_1) + \cos^2(\theta_2)$$

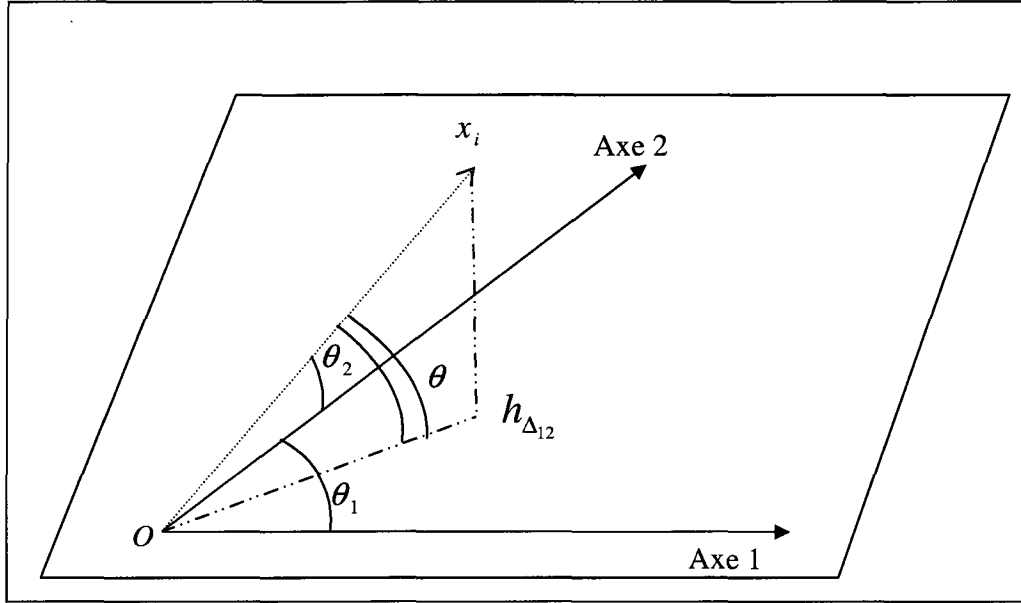


Figure 2 : Angle entre une observation et sa projection dans un plan

Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension k quelconque. Si le carré du cosinus de l'angle entre x_i et l'axe (ou le plan, ou le sous-espace) est proche de 1, alors on pourra dire que l'observation x_i est bien représentée par sa projection sur l'axe (ou le plan, ou le sous-espace) et si deux observations sont bien représentées en projection sur un axe (ou un plan, ou un sous-espace) et ont des projections proches, alors on pourra en conclure que ces deux observations sont proches dans l'espace. Le carré du cosinus de l'angle θ_{ik} entre x_i et un axe Δ_k de vecteur directeur unitaire μ_k est égal à :

$$\cos^2(\theta_{ik}) = \frac{\langle x_i, \mu_k \rangle^2}{\|x_i\|^2} = \frac{\mu_k^T x_i x_i^T \mu_k}{x_i^T x_i} = \frac{\left(\sum_{j=1}^p x_{ij} \mu_{kj} \right)^2}{\sum_{j=1}^p x_{ij}^2}$$

$$\cos^2(\theta_{ik}) = \frac{x_{ik}^2}{\sum_{j=1}^p x_{ij}^2}$$

On peut calculer le carré du cosinus de l'angle θ_{ikl} entre x_i et le plan engendré par deux axes $\Delta_k \oplus \Delta_l$

$$\cos^2(\theta_{ikl}) = \cos^2(\theta_{ik}) + \cos^2(\theta_{il})$$

Si après l'étude des pourcentages d'inertie expliquée par les sous-espaces successifs engendrés par les nouveaux axes, on a décidé de ne retenir qu'un sous-espace de dimension $d < p$, on pourra calculer la qualité de la représentation d'une observation x_i en calculant le carré du cosinus de l'angle de x_i avec ce sous-espace de \mathcal{R}^p .

1.2.10 Représentation des variables

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celui des observations. Les raisonnements se font dans \mathcal{R}^n au lieu de \mathcal{R}^p . Mais dans l'ACP, au delà de la symétrie formelle entre les observations et les variables, on peut utiliser la dissymétrie liée à la sémantique : les variables n'ont pas la même signification que les observations. On peut alors faire le raisonnement suivant : on a représenté les observations dans l'espace des anciennes variables, et on a fait un changement de base dans

cet espace. Les nouveaux axes sont des combinaisons linéaires des anciens axes et peuvent donc être considérés comme de nouvelles variables combinaisons linéaires des anciennes.

On notera $Z_1, Z_2, \dots, Z_k, \dots, Z_p$ les composantes principales où Z_k représente la nouvelle variable correspondant à l'axe Δ_k :

$$Z_k = \sum_{j=1}^p \mu_{kj} v_j = X \mu_k$$

Il est alors intéressant de voir comment les anciennes variables sont liées aux nouvelles; pour ce faire, on calcule les corrélations des anciennes variables avec les nouvelles. La représentation des anciennes variables se fera en prenant comme coordonnées des anciennes variables leurs coefficients de corrélation avec les nouvelles variables. On obtient alors ce que l'on appelle communément le cercle des corrélations, dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et +1, les représentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1, si on fait la représentation sur un plan.

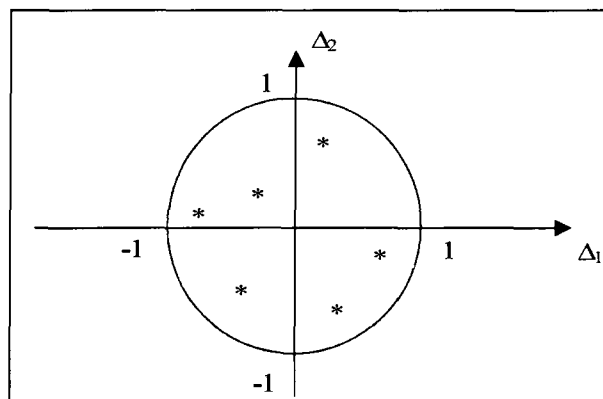


Figure 3: Cercle des corrélations pour les axes Δ_1 et Δ_2

Les variances des composantes principales sont égales à :

$$Var (Z_k) = \mu_k^T X^T D X \mu_k$$

Dans le cas particulier où les poids $p_1 = p_2 = \dots = p_n = \frac{1}{n}$ et que les données sont centrées, les variances s'écrivent de la façon suivante :

$$Var (Z_k) = \frac{1}{n} \mu_k^T X^T X \mu_k = \mu_k^T \Sigma \mu_k = \lambda_k$$

Dans ce cas, les covariances et coefficients de corrélation empiriques des composantes principales entre elles ou avec les variables de départ sont égales à :

$$Cov(Z_k, v_j) = \frac{1}{n} \mu_k^T X^T v_j = \frac{1}{n} \mu_k^T X^T X \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \mu_k^T \Sigma \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \lambda_k \mu_{kj}$$

et :

$$Cor(Z_k, v_j) = \sqrt{\lambda_k} \frac{\mu_{kj}}{\sqrt{Var(v_j)}}$$

où μ_{kj} et la $j^{ième}$ coordonnée du vecteur directeur unitaire μ_k de Δ_k .

La matrice de variance-covariance des composantes principales Σ_Z est égale à :

$$\Sigma_Z = A^T X^T X A = A^T \Sigma A = M$$

où A est la matrice du changement de base qui contient les vecteurs propres orthogonaux et de norme 1 et M est la matrice diagonale des valeurs propres de Σ :

$$M = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

La matrice des covariances entre les composantes principales et les anciennes variables peut s'écrire comme:

$$\text{Cov}(Z, V) = X^T X A = \Sigma A = A M$$

1.2.11 Qualité de la représentation des variables

Pour les mêmes raisons qui ont poussé à se préoccuper de la qualité de la représentation des observations, il faut se préoccuper de la qualité de la représentation des variables sur un axe, un plan ou un sous-espace. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. En effet, le coefficient de corrélation empirique entre une ancienne variable v_j et une nouvelle variable Z_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point x_j représentant la variable v_j avec cet axe.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1.

1.2.12 Études des liaisons entre les variables

À partir du graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les unes par rapport aux autres en termes de corrélations. Pour illustrer ce type d'interprétation, on utilise un exemple où certaines variables démographiques de la population des États-Unis sont projetées dans le cercle des corrélations. Deux points très proches du cercle des corrélations, donc bien représentées dans le plan, seront fortement corrélés positivement comme le taux de mortalité et la population de 65 ans et plus dans la figure 4¹. Ceci s'explique car, plus la population est âgée et plus le nombre de décès est grand. Si deux variables sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement comme la population de 65 ans et plus et le taux de naissance car, plus la population est âgée et moins il y aura de naissance.

¹ Les données proviennent du US Census Bureau. Le fichier original peut-être obtenu sur http://eire.census.gov/popest/states_dataset.csv.

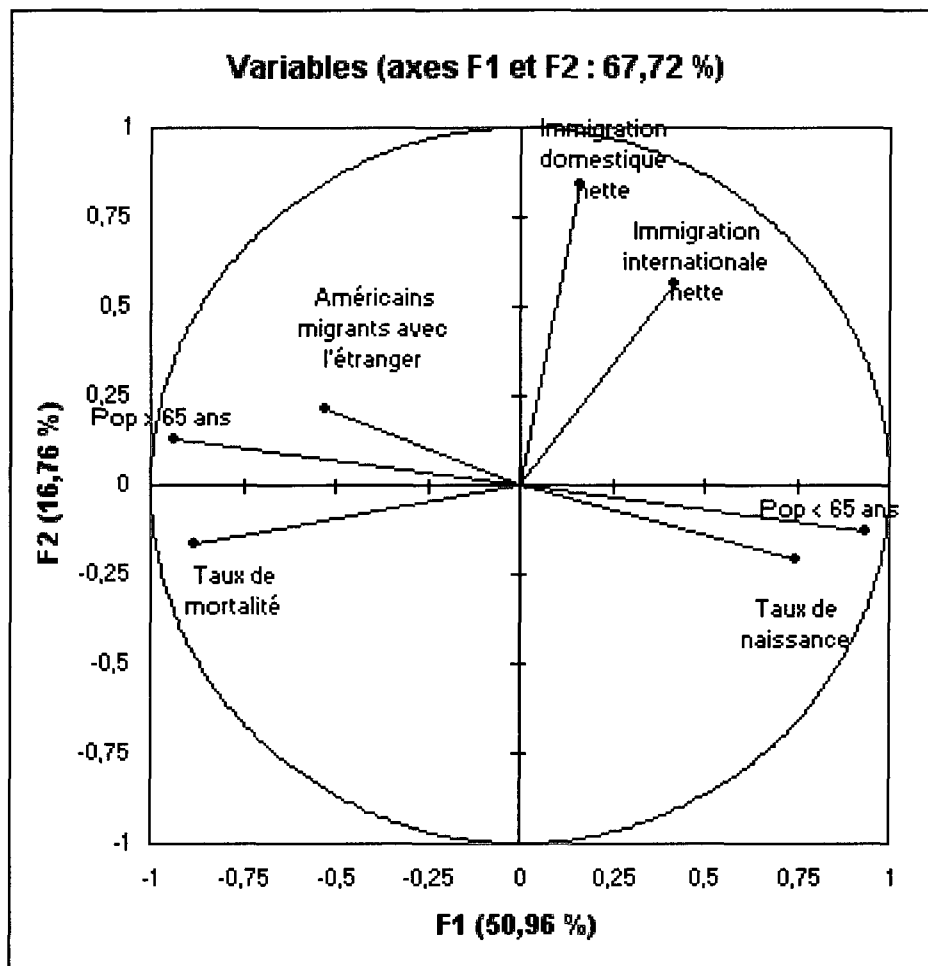


Figure 4: Exemple d'un cercle des corrélations

Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles comme le taux de naissance et l'immigration domestique nette (voir figure 4). Cela est logique car, l'immigration domestique représente la migration interne, soit les Américains qui ne résident plus dans leur État de naissance et qu'il n'y a pas de grande variation entre la fécondité de la population locale et des immigrants internes, ce qui explique la corrélation pratiquement nulle.

Il faut, pour interpréter correctement ces graphiques des cercles de corrélation, se souvenir qu'un coefficient de corrélation est une mesure de liaison linéaire entre deux variables; il peut arriver que deux variables très fortement liées aient un coefficient de corrélation nul ou très faible, si leur liaison n'est pas linéaire.

1.2.13 Observations et variables supplémentaires

Il arrive que l'on veuille faire apparaître dans les représentations graphiques certaines observations sans qu'elles interviennent dans la détermination des axes. Cela peut être le cas de nouvelles observations que l'on veut simplement positionner par rapport aux autres sans que les positions de celles-ci soient influencées par les nouvelles. On dit de ces dernières, qu'elles sont des observations supplémentaires.

Il en est de même pour les variables. On peut, par exemple, vouloir représenter une variable supplémentaire dans l'espace des p variables choisies pour faire l'ACP, afin de mieux comprendre comment cette variable est liée aux anciennes. Cependant on ne souhaite pas qu'elle intervienne dans la construction des axes car ses liaisons avec les p variables de départ fausseraient la représentation si elle faisait partie intégrante de l'ACP.

Pour représenter une observation supplémentaire, il suffit d'exprimer les coordonnées de cette observation dans la nouvelle base des axes principaux. Pour une variable supplémentaire, il suffit de calculer ses coefficients de corrélation empiriques avec les composantes principales.

1.2.14 Exemple

Pour faciliter la compréhension de l'analyse en composantes principales, cette sous-section présente un exemple tiré de [Bouroche *et al.* 1980]. Il s'agit de l'étude des différents postes du budget de l'état français de 1872 à 1971. Les valeurs sont données en pourcentage du budget global pour éliminer l'effet de l'évolution de la valeur nominale du franc au cours du temps. Les intitulés complets des variables sont :

- AN : Année (irrégulièrement espacées de 1872 à 1971),
- PVP : Pouvoirs publics,
- AGR : Agriculture,
- CMI : Commerce et industrie,
- TRA : Travail,
- LOG : Logement,
- EDU : éducation,
- ACS : Action sociale,
- ANC : Anciens combattants,
- DEF : Défense,
- DET : Remboursement de la dette,
- DIV : Divers.

Le tableau 1 correspond à la matrice X des données.

Tableau 1: Présentation des données

Budgets de l'état de 1872 à 1971 : ACP simple												
Données brutes												
OBS	AN	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1	1872	18,0	0,5	0,1	6,7	0,5	2,1	2,0	0,0	26,4	41,5	2,1
2	1880	14,1	0,8	0,1	15,3	1,9	3,7	0,5	0,0	29,8	31,3	2,5
3	1890	13,6	0,7	0,7	6,8	0,6	7,1	0,7	0,0	33,8	34,4	1,7
4	1900	14,3	1,7	1,7	6,9	1,2	7,4	0,8	0,0	37,7	26,2	2,2
5	1903	10,3	1,5	0,4	9,3	0,6	8,5	0,9	0,0	38,4	27,2	3,0
6	1906	13,4	1,4	0,5	8,1	0,7	8,6	1,8	0,0	38,5	25,3	1,9
7	1909	13,5	1,1	0,5	9,0	0,6	9,0	3,4	0,0	36,8	23,5	2,6
8	1912	12,9	1,4	0,3	9,4	0,6	9,3	4,3	0,0	41,1	19,4	1,3
9	1920	12,3	0,3	0,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	0,2
10	1923	7,6	1,2	3,2	5,1	0,6	5,6	1,8	10,0	29,0	35,0	0,9
11	1926	10,5	0,3	0,4	4,5	1,8	6,6	2,1	10,1	19,9	41,6	2,3
12	1929	10,0	0,6	0,6	9,0	1,0	8,1	3,2	11,8	28,0	25,8	2,0
13	1932	10,6	0,8	0,3	8,9	3,0	10,0	6,4	13,4	27,4	19,2	0,0
14	1935	8,8	2,6	1,4	7,8	1,4	12,4	6,2	11,3	29,3	18,5	0,4
15	1938	10,1	1,1	1,2	5,9	1,4	9,5	6,0	5,9	40,7	18,2	0,0
16	1947	15,6	1,6	10,1	11,4	7,6	8,8	4,8	3,4	32,2	4,6	0,0
17	1950	11,2	1,3	16,5	12,4	15,8	8,1	4,9	3,4	20,7	4,2	1,5
18	1953	12,9	1,5	7,0	7,9	12,1	8,1	5,3	3,9	36,1	5,2	0,0
19	1956	10,9	5,3	9,7	7,6	9,6	9,4	8,5	4,6	28,2	6,2	0,0
20	1959	13,1	4,4	7,3	5,7	9,8	12,5	8,0	5,0	26,7	7,5	0,0
21	1962	12,8	4,7	7,5	6,6	6,8	15,7	9,7	5,3	24,5	6,4	0,1
22	1965	12,4	4,3	8,4	9,1	6,0	19,5	10,6	4,7	19,8	3,5	1,8
23	1968	11,4	6,0	9,5	5,9	5,0	21,1	10,7	4,2	20,0	4,4	1,9
24	1971	12,8	2,8	7,1	8,5	4,0	23,8	11,3	3,7	18,8	7,2	0,0

Les observations sont les années alors que les variables sont les pourcentages de budgets alloués aux différents postes. Chaque année est représentée par un point situé dans un espace de dimension 11 dont les coordonnées sont les pourcentages des différents postes dans le budget national de cette année-là. En fait, le nuage des observations est compris dans un sous-espace de dimension 10 puisque la somme des différents pourcentages pour une année donnée vaut toujours 100%. Chaque poste budgétaire est représenté par un point situé dans

un espace de dimension 24 dont les coordonnées sont les pourcentages accordés à ce poste aux différentes années.

Le tableau 2 présente les centres de gravités et les écart-types pour chacune des variables.

Tableau 2: Centres de gravités et écarts-types

Moyenne et écart-type											
	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
Moyenne	12,2	2,0	3,9	8,3	4,0	9,9	4,8	4,3	30,3	19,1	1,2
Écart-type	2,2	1,6	4,5	2,5	4,2	5,2	3,4	4,2	7,3	12,2	1,0

Après le changement d'origine, la coordonnée de l'année 1872 sur l'axe défini par la variable PVP devient :

$$18,0 - 12,2 = 5,8$$

Cette valeur est également la nouvelle coordonnée de la variable PVP sur l'axe défini par l'année 1872. Cela est aussi vrai pour chacune des variables et pour chacune des années représentées dans le tableau 1.

Le tableau 3 présente la matrice de covariance Σ pour les 11 variables qui sont étudiées. Dans ce cas la variance totale est de 310,57 et elle nous est donnée par :

$$trace(\Sigma) = 310,57$$

Tableau 3: Matrice de variance-covariance

Matrice de covariance											
	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
PVP	4,80	-0,30	0,00	1,26	0,33	-1,72	-0,98	-6,25	1,62	0,90	0,34
AGR	-0,30	2,71	4,43	-1,12	3,00	6,29	4,52	0,30	-5,39	-13,95	-0,47
CMI	0,00	4,43	20,15	1,03	16,73	10,95	9,51	0,42	-17,60	-44,02	-1,60
TRA	1,26	-1,12	1,03	6,09	1,71	-2,75	-1,71	-3,21	2,85	-4,46	0,29
LOG	0,33	3,00	16,73	1,71	17,49	5,08	6,95	0,78	-11,57	-38,65	-1,88
EDU	-1,72	6,29	10,95	-2,75	5,08	27,28	15,58	3,41	-20,01	-42,69	-1,33
ACS	-0,98	4,52	9,51	-1,71	6,95	15,58	11,62	4,08	-14,13	-33,59	-1,85
ANC	-6,25	0,30	0,42	-3,21	0,78	3,41	4,08	17,26	-12,66	-2,50	-1,61
DEF	1,62	-5,39	-17,60	2,85	-11,57	-20,01	-14,13	-12,66	53,43	23,32	0,15
DET	0,90	-13,95	-44,02	-4,46	-38,65	-42,69	-33,59	-2,50	23,32	148,69	6,93
DIV	0,34	-0,47	-1,60	0,29	-1,88	-1,33	-1,85	-1,61	0,15	6,93	1,05

On peut remarquer que les variances sont très hétérogènes : $\text{Var}(\text{DET}) = 148,69$, $\text{Var}(\text{DIV}) = 1,05$. La contribution du remboursement de la dette à l'inertie totale est donc beaucoup plus forte; ce poste influera plus sur les résultats de l'analyse que le poste "Divers".

Le tableau 4 nous donne les inerties (valeurs propres) associées aux différents axes. On peut vérifier que leur somme est bien égale à l'inertie totale :

$$\sum_i \lambda_i = \sum_i I_{\Delta_i} = 310,57$$

La colonne proportion cumulée nous indique que les axes Δ_1 , Δ_2 et Δ_3 portent à eux trois 90,96 pourcent de l'inertie totale. On utilise souvent la colonne différence pour choisir le nombre d'axes à conserver dans l'étude. Ici on trouve une forte chute entre les axes Δ_4 et Δ_5 ce qui, joint à la forte part d'inertie portée par Δ_1 , Δ_2 et Δ_3 , suggère de conserver les 3 ou les 4 premiers axes dans l'analyse.

Tableau 4: Valeurs propres de la matrice de variance-covariance

Valeurs propre de la matrice de covariance				
	Valeurs propres	Différence	Proportion	Proportion cumulée
AXE1	204,717	148,896	0,659161	0,65916
AXE2	55,822	33,853	0,179738	0,8389
AXE3	21,969	5,566	0,070736	0,90963
AXE4	16,402	10,88	0,052813	0,96245
AXE5	5,522	2,444	0,01778	0,98023
AXE6	3,078	1,966	0,00991	0,99014
AXE7	1,111	0,039	0,003578	0,99372
AXE8	1,072	0,44	0,003452	0,99717
AXE9	0,632	0,384	0,002034	0,9992
AXE10	0,248	0,248	0,000798	1
AXE11	0		0	1

Le tableau 5 donne les coordonnées y_{ij} des différentes années dans le nouveau système d'axes et la figure 5 montre la projection du nuage dans le premier plan principal (plan 1-2).

Budgets de l'etat de 1872 a 1971 : ACP simple
Projection dans le plan 1-2

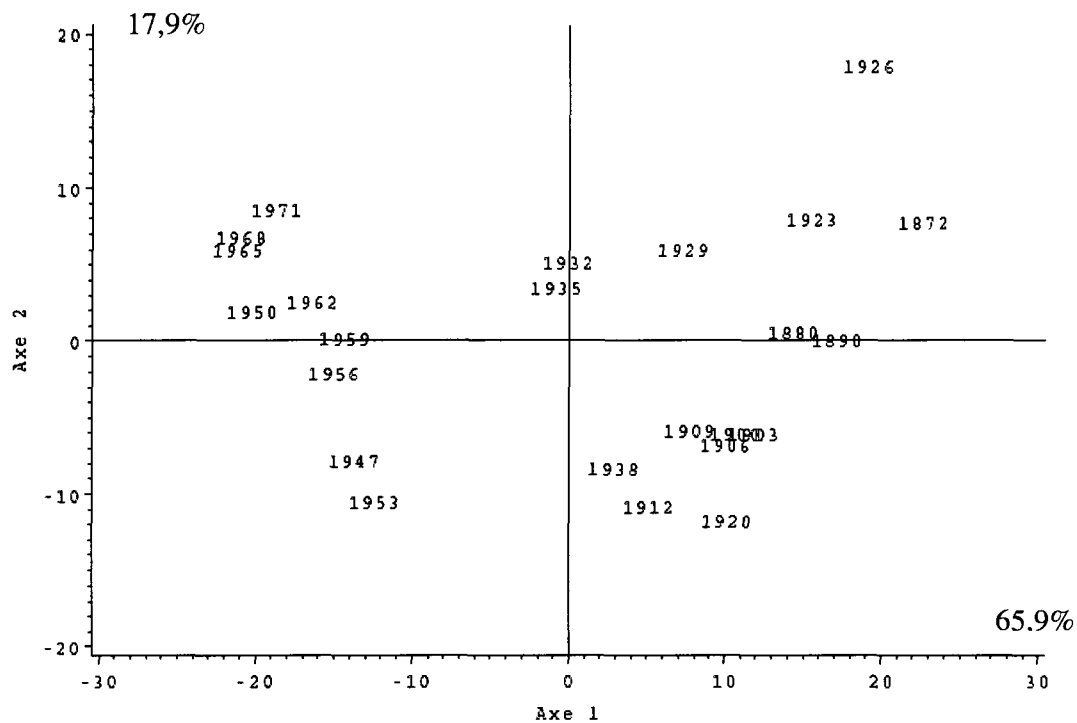


Figure 5 : Projection du nuage dans le premier plan principal (plan 1-2)

Tableau 5: Composantes principales pour chacune des années

Coordonnées factorielles des observations												
OBS	AN	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	AXE11
1	1872	-22,6	7,8	-6,3	5,8	1,7	-4,3	-0,5	0,9	-0,6	-0,6	0,0
2	1880	-14,4	0,7	-5,9	3,7	-6,3	0,0	-1,6	-0,9	0,2	0,5	0,0
3	1890	-17,1	0,2	-0,9	4,0	2,1	0,7	1,0	-0,2	-0,3	0,6	0,0
4	1900	-10,7	-6,0	0,1	2,9	2,1	0,3	0,9	0,9	1,4	0,4	0,0
5	1903	-11,8	-6,0	1,8	2,2	-0,1	3,6	-0,3	-1,1	0,7	-0,1	0,0
6	1906	-10,0	-6,8	1,9	2,9	0,7	0,6	0,6	0,0	0,6	0,3	0,0
7	1909	-7,7	-5,8	2,1	3,4	-0,6	0,0	0,0	-0,1	0,3	-1,0	0,0
8	1912	-5,1	-10,8	4,0	2,1	-0,6	-0,1	-0,4	-0,1	-0,6	-0,5	0,0
9	1920	-10,1	-11,7	-0,6	-2,1	-2,3	-0,2	-0,6	-0,3	-1,1	0,5	0,0
10	1923	-15,5	8,0	0,0	-5,8	2,8	3,1	-0,4	1,5	-0,9	0,4	0,0
11	1926	-19,1	18,0	-2,1	-2,4	1,9	0,0	0,8	-1,1	0,2	-0,2	0,0
12	1929	-7,2	6,0	2,2	-6,0	-3,0	0,0	0,5	0,6	1,2	-0,5	0,0
13	1932	0,1	5,2	3,9	-7,4	-3,3	-2,6	0,5	-0,5	0,0	-0,1	0,0
14	1935	0,8	3,6	6,2	-5,4	-1,5	0,3	-0,5	0,1	0,3	0,5	0,0
15	1938	-2,9	-8,3	6,3	-4,0	2,2	-0,1	0,0	0,3	-1,5	-0,8	0,0
16	1947	13,8	-7,7	-4,8	-0,9	-1,8	-1,3	1,4	3,2	0,3	0,4	0,0
17	1950	20,3	2,0	-14,3	-1,9	-0,9	3,3	0,9	-0,3	-0,6	-0,6	0,0
18	1953	12,4	-10,4	-3,9	-4,2	2,2	-1,6	1,5	-2,1	0,4	-0,1	0,0
19	1956	15,0	-2,1	-3,4	-2,6	2,5	-0,2	-3,1	0,1	-0,4	0,1	0,0
20	1959	14,4	0,2	-1,3	-0,8	3,1	-2,1	-0,3	-1,3	0,7	0,7	0,0
21	1962	16,4	2,6	1,8	1,0	1,4	-1,5	-0,7	0,1	0,1	0,4	0,0
22	1965	21,2	6,0	2,6	3,8	-1,9	0,1	-0,4	0,0	0,5	-0,6	0,0
23	1968	21,0	6,9	4,4	4,6	1,6	1,5	-0,9	0,8	0,9	-0,2	0,0
24	1971	18,7	8,7	6,0	6,9	-1,8	0,3	1,8	-0,6	-1,9	0,5	0,0

Le tableau 6 donne les cosinus carrés $\cos^2(\alpha_{ik})$ pour les axes Δ_1 , Δ_2 , Δ_3 (COS2_1, COS2_2, COS2_3), pour le plan 1-2 (COS2_1_2) et pour l'espace 1-2-3 (COS2_1_2_3). La lecture de ce tableau démontre que la plupart des années sont bien représentées dans le plan 1-2. On peut cependant noter que les années 1932 et 1935 ne sont pas très bien représentées. Leur apparente proximité sur la figure 2 ne peut donc pas être directement interprétée. On peut d'ailleurs noter, en revenant au tableau 5 sur les axes principaux, que leurs coordonnées sur l'axe Δ_3 sont différentes (3.85 pour 1932 et 6.22 pour 1935).

Tableau 6: Qualité de la représentation

Qualité de la représentation des observations (cosinus carrés)						
OBS	AN	COS2_1	COS2_2	COS2_12	COS2_3	COS2_1_2_3
1	1872	0,76	0,09	0,86	0,06	0,91
2	1880	0,69	0,00	0,69	0,12	0,81
3	1890	0,93	0,00	0,93	0,00	0,93
4	1900	0,68	0,21	0,90	0,00	0,90
5	1903	0,70	0,18	0,88	0,02	0,90
6	1906	0,63	0,29	0,91	0,02	0,94
7	1909	0,54	0,31	0,84	0,04	0,88
8	1912	0,16	0,71	0,87	0,10	0,97
9	1920	0,41	0,55	0,95	0,00	0,95
10	1923	0,67	0,18	0,85	0,00	0,85
11	1926	0,52	0,46	0,98	0,01	0,98
12	1929	0,37	0,26	0,63	0,03	0,66
13	1932	0,00	0,24	0,24	0,13	0,37
14	1935	0,01	0,15	0,16	0,46	0,62
15	1938	0,06	0,49	0,55	0,28	0,83
16	1947	0,65	0,20	0,86	0,08	0,94
17	1950	0,65	0,01	0,65	0,32	0,97
18	1953	0,50	0,35	0,85	0,05	0,90
19	1956	0,85	0,02	0,87	0,04	0,92
20	1959	0,91	0,00	0,91	0,01	0,92
21	1962	0,94	0,02	0,97	0,01	0,98
22	1965	0,88	0,07	0,95	0,01	0,96
23	1968	0,82	0,09	0,91	0,04	0,95
24	1971	0,67	0,14	0,82	0,07	0,89

Les coordonnées des nouvelles variables dans l'espace des observations se lisent en colonne dans le tableau 5. Les corrélations des nouvelles variables avec les anciennes sont données dans le tableau 7.

Tableau 7: Corrélations des nouvelles variables avec les anciennes

Coefficient de corrélation de Pearson (Rho)										
	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10
PVP	-0,07	-0,24	-0,35	0,66	-0,04	-0,58	0,15	0,14	0,07	0,02
	0,76	0,27	0,10	0,00	0,86	0,00	0,48	0,51	0,76	0,93
AGR	0,75	0,14	0,21	0,20	0,28	0,00	-0,44	0,03	0,19	0,14
	0,00	0,51	0,32	0,35	0,18	0,98	0,03	0,89	0,39	0,50
CMI	0,86	0,09	-0,42	0,02	0,14	0,17	0,03	0,13	-0,04	0,00
	0,00	0,69	0,04	0,93	0,50	0,43	0,91	0,53	0,84	1,00
TRA	0,07	-0,35	-0,39	0,10	-0,83	0,11	-0,06	-0,02	-0,05	0,03
	0,75	0,09	0,06	0,65	0,00	0,60	0,79	0,94	0,81	0,90
LOG	0,78	-0,06	-0,56	-0,17	0,13	-0,03	0,04	-0,17	0,00	0,01
	0,00	0,78	0,00	0,42	0,55	0,91	0,84	0,42	1,00	0,95
EDU	0,75	0,27	0,52	0,29	0,00	0,09	0,08	-0,04	0,00	0,02
	0,00	0,20	0,01	0,17	1,00	0,66	0,70	0,86	0,99	0,92
ACS	0,88	0,24	0,32	0,05	0,06	-0,17	-0,12	-0,02	-0,13	-0,06
	0,00	0,26	0,12	0,81	0,77	0,44	0,57	0,91	0,53	0,77
ANC	0,13	0,54	0,26	-0,78	-0,07	-0,06	0,04	0,05	0,03	0,01
	0,53	0,01	0,21	0,00	0,74	0,77	0,84	0,83	0,89	0,98
DEF	-0,47	-0,86	0,16	-0,10	0,05	0,02	0,01	0,00	-0,01	0,00
	0,02	0,00	0,45	0,65	0,81	0,93	0,96	0,99	0,97	1,00
DET	-0,97	0,22	-0,03	0,03	0,03	0,01	0,00	-0,01	-0,01	0,00
	0,00	0,31	0,90	0,87	0,90	0,94	0,99	0,98	0,95	0,99
DIV	-0,51	0,17	-0,13	0,46	-0,12	0,36	-0,04	-0,06	0,45	-0,34
	0,01	0,42	0,53	0,02	0,57	0,08	0,84	0,79	0,03	0,10

La lecture du tableau des corrélations entre les anciennes et les nouvelles variables montre une forte corrélation négative entre le premier axe et la dette ainsi que de fortes corrélations positives entre ce même axe et le commerce, le logement et l'agriculture. L'opposition entre ces 2 groupes de variables se retrouve dans le cercle des corrélations dans le plan 1-2 (figure 6). Cette opposition entre DET et le groupe {EDU, ACS, AGR, CMI, LOG} (qui était déjà visible dans la matrice de covariance) peut fournir une interprétation de l'axe Δ_1 à des personnes ayant une bonne connaissance de la situation politique et économique en France. L'axe Δ_2 est lui très lié à la variable DEF : ce poste budgétaire semble avoir une évolution différente des autres postes. La proximité des variables EDU, ACS, AGR, CMI et LOG sur la figure 3 permet de conclure

qu'il existe un lien fort entre ces variables mais cette interprétation ne vaut que parce ces variables sont bien représentées dans le plan 1-2 (les points correspondants sont proches du cercle). La proximité des variables PVP et TRA ne peut pas être interprétée car ces variables sont inadéquatement représentées puisque les points correspondants sont éloignés du cercle. La lecture de la matrice de covariance montre d'ailleurs que le lien est assez faible : $\text{Var}(\text{PVP}) = 4.801$, $\text{Var}(\text{TRA}) = 6.0899$, $\text{Cov}(\text{PVP}, \text{TRA}) = 1.2585$.

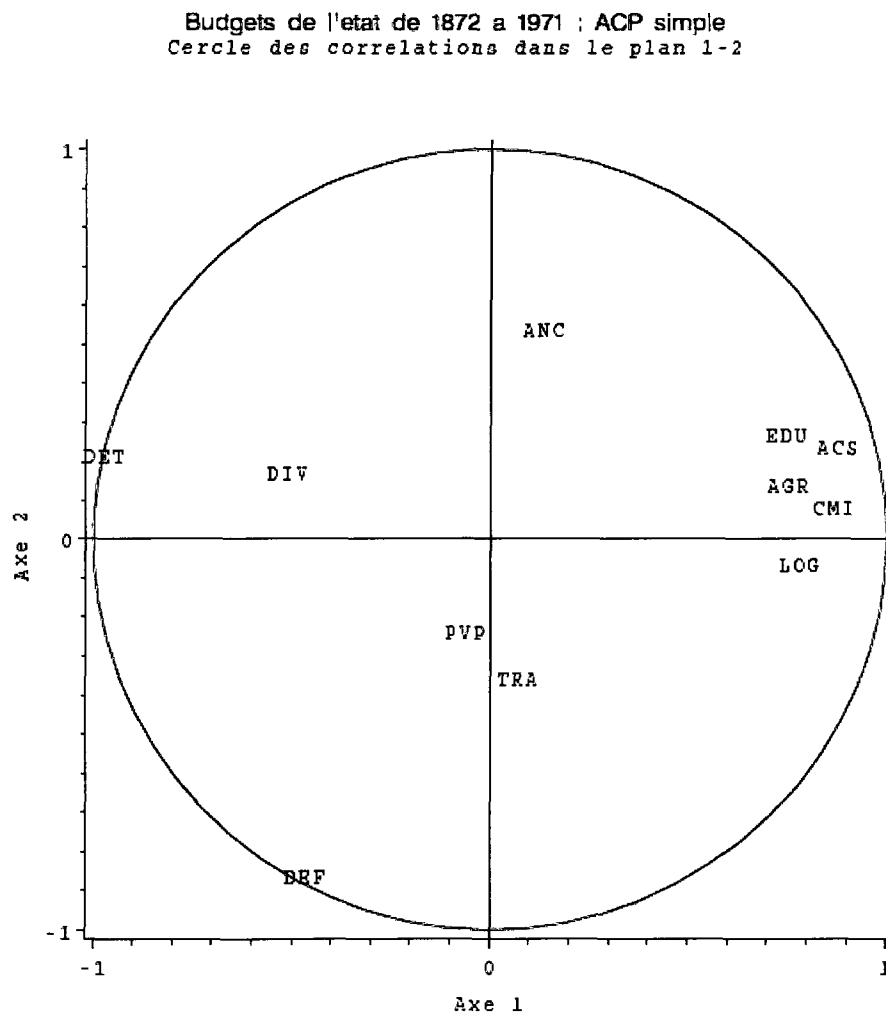


Figure 6: Cercle des corrélations

1.3 Analyse factorielle discriminante (AFD)

1.3.1 Introduction

L'analyse factorielle discriminante est une méthode d'analyse de données qui se situe à mi-chemin entre une technique descriptive et une technique décisionnelle. Elle a été introduite par Ronald A. Fisher en 1936 [Fisher 1936]. L'AFD s'applique à des données quantitatives d'une population ou d'un échantillon de n observations qui sont, *a priori*, divisées en k catégories. Chacune des observations est décrites par p variables. On peut distinguer le plus souvent deux grandes approches.

Premièrement, l'analyse discriminante peut être une technique descriptive. On parle dans ce cas d'analyse factorielle discriminante (ou analyse discriminante descriptive). L'objectif est de proposer un nouveau dispositif de représentation, des variables latentes constituées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner au maximum les k groupes dans lesquels sont classées les observations. En ce sens, elle se rapproche de l'analyse factorielle car elle sert à proposer une représentation graphique dans un espace réduit, et plus particulièrement de l'analyse en composantes principales effectuée sur les centres de gravité des nuages de points de chacune des catégories avec une métrique spécifique. On parle aussi d'analyse canonique discriminante, surtout dans les logiciels et manuels anglo-saxons.

Deuxièmement, l'analyse discriminante peut être prédictive. Il s'agit dans ce cas de construire une fonction de classement, où règle d'affectation, qui sert à prédire le groupe d'appartenance d'une observation à partir des valeurs prises par les variables prédictives. En ce sens, cette technique se rapproche des techniques supervisées en apprentissage automatique telles que les arbres de décision, les réseaux de neurones, ... Elle repose sur un cadre probabiliste. Le plus connu est sans doute l'hypothèse de distribution multinormale (loi normale).

Additionnée à l'hypothèse d'homoscédasticité, on suppose que les nuages de points de chacun des groupes ont la même forme, ce qui mène à l'analyse discriminante linéaire. Elle est particulièrement séduisante dans la pratique car la fonction de classement s'exprime comme une combinaison linéaire des variables prédictives, facile à analyser ainsi qu'à interpréter. Cette technique est, avec la régression logistique, particulièrement utilisée dans le « scoring », quand on désire par exemple caractériser l'appétence – la propension à acheter – d'un client face à un nouveau produit. La distinction entre ces deux approches n'est pas aussi tranchée. Il est par exemple envisageable de dériver des règles géométriques d'affectation à partir de l'analyse discriminante.

Voici un cas où l'analyse discriminante s'avère efficace. La figure 7 représente deux classes C_1 et C_2 décrites par deux variables x_1 et x_2 . Aucune de ces deux variables prise séparément ne peut séparer efficacement les deux classes en raison du chevauchement important des projections des classes sur les axes représentant ces variables.

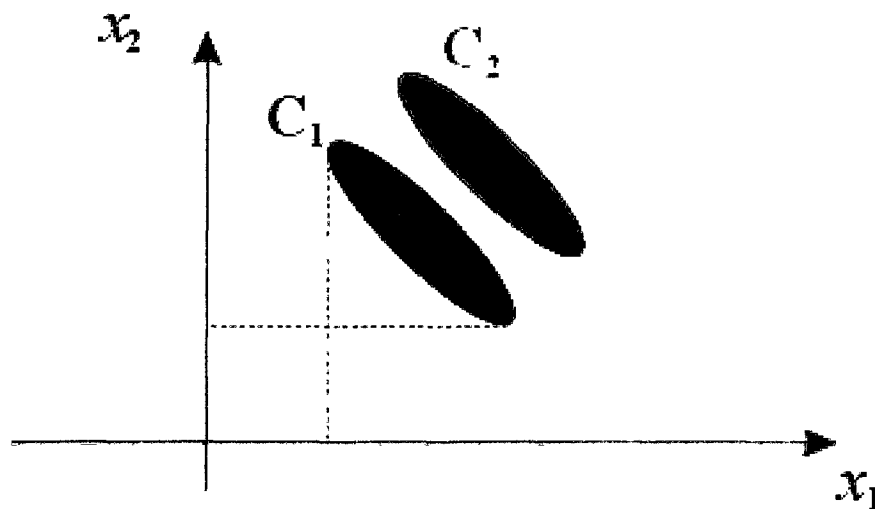


Figure 7 : Illustration des données dans le cas de deux classes et de deux variables

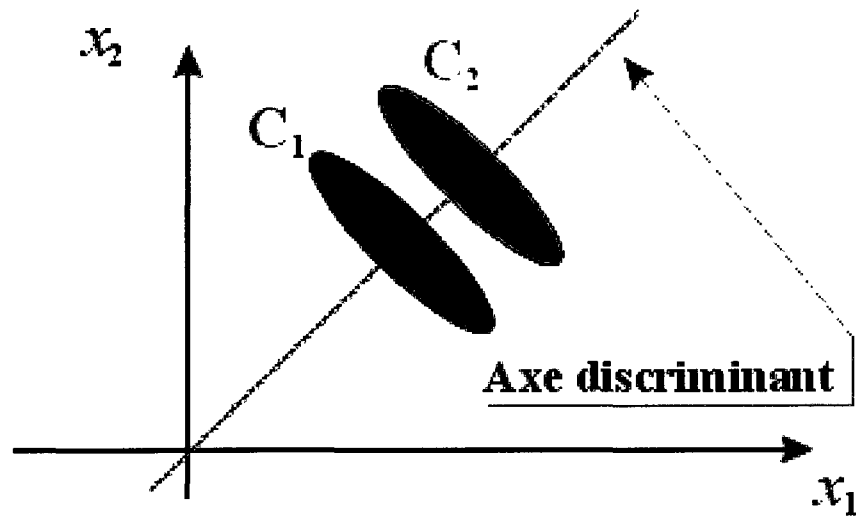


Figure 8 : Illustration de l'axe discriminant dans le cas de deux classes et de deux variables

Par contre, la figure 8 montre que la nouvelle direction définie par $x_1 = x_2$ est totalement discriminante car les projections des classes sur cette direction sont parfaitement séparées.

1.3.2 Données

Les données, la matrice X , se présentent de la même façon que dans l'ACP, soit :

$$X = \{x_{ij}\}_{i:1,2,\dots,n, j:1,2,\dots,p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \quad \text{où } x_{ij} \in \mathbb{R}$$

et

$$X = [x_1, x_2, \dots, x_p]$$

De plus, on suppose que les n observations sont réparties, *a priori*, en k catégories disjointes.

Cela revient à créer une variable qualitative T , à k modalités $\{T_1, \dots, T_k\}$, jouant le rôle de variable à expliquer.

La situation est similaire à celle de la régression linéaire multiple mais, comme la variable à expliquer est qualitative, on parvient à une méthode très différente qui est exposée dans la section suivante.

Dans l'analyse factorielle discriminante, les données des n observations sont affectées des poids $p_i > 0$ que l'on pose :

$$D = \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{bmatrix}$$

où

$$\sum_{i=1}^n p_i = 1$$

La variable T engendre une partition $\{T_q; q = 1, \dots, k\}$ de l'ensemble des observations dont chaque élément est d'effectif n_q .

On désigne par Q la matrice de dimension $n \times k$ des indicatrices des modalités de la variable T . Le terme général de cette matrice, q_{ij} , est égal à 1 si x_i appartient à la classe T_j et 0 sinon, où $i : 1, 2, \dots, n$ et $j : 1, 2, \dots, k$.

Si l'on pose :

$$\overline{p}_q = \sum_{i: x_i \in \Omega_q} p_i$$

on en déduit \overline{D} la matrice diagonale des poids des éléments de la partition engendrée par T tel que :

$$\overline{D} = Q^T D Q = \begin{bmatrix} \overline{p}_1 & & 0 \\ & \ddots & \\ 0 & & \overline{p}_k \end{bmatrix}$$

On suppose par ailleurs que les variables sont centrées sur l'ensemble des observations.

1.3.3 Notations

Les coordonnées, g_j , du centre de gravité, G confondu avec l'origine, pour les p variables de l'échantillon total E est donné par :

$$g_j = \sum_{i=1}^n p_i x_{ij} \quad j = 1, 2, \dots, p$$

Sans perte de généralité, on suppose que :

$$p_i = \frac{1}{n} \quad \forall i : 1, 2, \dots, n$$

Les coordonnées, g_j^q , du centre de gravité, G^q , pour les k catégories et les p variables sont données par :

$$g_j^q = \frac{1}{n_q} \sum_{x_i \in E_q} x_{ij} \quad j = 1, 2, \dots, p \text{ et } q = 1, 2, \dots, k$$

où n_q est le nombre d'observation appartenant à la catégorie q . On a donc :

$$\sum_{q=1}^k n_q = n$$

Un concept mathématique important en analyse discriminante est celui de la covariance. La covariance entre deux variables j et l , noté $\text{cov}(j, l)$, est définie comme :

$$\text{cov}(j, l) = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)$$

que l'on peut aussi écrire comme :

$$\text{cov}(j, l) = \frac{1}{n} \sum x_{ij} x_{il} - \bar{x}_j \bar{x}_l$$

On peut déterminer la covariance entre deux variables descriptives à l'intérieur de l'une des k catégories :

$$\text{cov}_q(j, l) = \frac{1}{n_q} \sum_{x_i \in T_q} (x_{ij} - g_j^q)(x_{il} - g_l^q) \quad q = 1, 2, \dots, k$$

La $\text{cov}(j, l)$ peut s'écrire en utilisant la même notion de catégorie :

$$\text{cov}(j, l) = \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} (x_{ij} - g_j)(x_{il} - g_l) \right]$$

Comme en analyse de variance, nous pouvons aussi déterminer la covariance intra-classe, $\text{cov}_{\text{intra}}(j, l)$ et la covariance interclasse, $\text{cov}_{\text{inter}}(j, l)$, qui se définissent comme suit :

$$\begin{aligned} \text{cov}_{\text{intra}}(j, l) &= \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} (x_{ij} - g_j^q)(x_{il} - g_l^q) \right] = \sum_{q=1}^k \frac{n_q}{n} \text{Cov}_q(j, l) \\ \text{cov}_{\text{inter}}(j, l) &= \sum_{q=1}^k \frac{n_q}{n} (g_j^q - g_j)(g_l^q - g_l) \end{aligned}$$

On en déduit donc que :

$$\text{cov}(j, l) = \text{cov}_{\text{intra}}(j, l) + \text{cov}_{\text{inter}}(j, l)$$

On peut démontrer l'équation précédente si l'on reprend la formulation suivant pour $\text{cov}(j, l)$:

$$\text{cov}(j, l) = \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} (x_{ij} - g_j)(x_{il} - g_l) \right]$$

En effet, on peut réécrire les termes de la façon suivante :

$$(x_{ij} - g_j) = (x_{ij} - g_j^q) + (g_j^q - g_j)$$

Ce qui nous donne :

$$\text{cov}(j,l) = \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} ((x_{ij} - g_j^q) + (g_j^q - g_j))((x_{il} - g_l^q) + (g_l^q - g_l)) \right]$$

et comme

$$\sum_{x_i \in T_q} (x_{ij} - g_j^q)(g_l^q - g_l) = \sum_{x_i \in T_q} (g_j^q - g_j)(x_{il} - g_l^q) = 0$$

on remarque que deux des quatre termes du produit croisé sont non nuls. Donc, cela revient à :

$$\text{cov}(j,l) = \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} (x_{ij} - g_j^q)(x_{il} - g_l^q) + (g_j^q - g_j)(g_l^q - g_l) \right]$$

$$\text{cov}(j,l) = \frac{1}{n} \sum_{q=1}^k \left[\sum_{x_i \in T_q} (x_{ij} - g_j^q)(x_{il} - g_l^q) \right] + \sum_{q=1}^k \frac{n_q}{n} (g_j^q - g_j)(g_l^q - g_l)$$

$$\text{cov}(j,l) = \text{cov}_{\text{intra}}(j,l) + \text{cov}_{\text{inter}}(j,l)$$

Avant de poursuivre, définissons la matrice S comme la matrice symétrique des covariances totales de terme général $\text{cov}(j,l)$, la matrice S_r comme la matrice symétrique des covariances intra-classes de terme général $\text{cov}_{\text{intra}}(j,l)$ et finalement la matrice S_e comme la matrice symétrique des covariances interclasses de terme général $\text{cov}_{\text{inter}}(j,l)$. Notons que :

$$S = S_e + S_r$$

1.3.4 Modèle

Dans l'espace des observations, le principe consiste à projeter ces observations selon une direction permettant de mettre le plus en évidence les différents groupes possibles. À cette fin, il convient de déterminer une direction Δ dans \mathfrak{R}^p selon laquelle la variance interclasse des projections des observations sur Δ est la plus grande possible et que la variance intra-classe de ces mêmes projections soit la plus petite possible.

En analyse en composantes principales, on ne dispose que d'une observation x_i pour chaque effet à estimer. Par contre, dans le cas de l'analyse discriminante, on peut considérer que les éléments d'une même classe T_q , $q: 1, 2, \dots, k$ sont des observations répétées n_q fois du même effet $z_q = \alpha_q T_q$, $\alpha_q \in \mathfrak{R}$, pondéré par les

$\overline{p_q} = \frac{n_q}{n}$ sous l'hypothèse que la variance intra-classe est nulle. Le modèle peut donc s'écrire comme un modèle linéaire conventionnel :

- $x_i; i = 1, \dots, n$ qui représente n vecteurs indépendants de $E \in \mathfrak{R}^p$
- $\forall k, \forall i \in T_q, x_i = z_1 + z_2 + \dots + z_q + \dots + z_k + \varepsilon_i = z_q + \varepsilon_i$
- L'espérance de ε_i est égal à 0 et la variance est égale à $w_{i,i}$
- W , de terme général $w_{i,j}$, la matrice de variance-covariance de ε_i ; est définie positive et son déterminant n'est pas égale à zéro

On suppose aussi qu'il existe un sous espace $A_m \subseteq \mathfrak{R}^p$ de dimension m de E tel que :

$$\forall q, z_q \in A_m; m \leq \min(p, k - 1)$$

Soit $\bar{z} = \sum_{q=1}^k \bar{p}_q z_q$, le modèle implique que $\bar{z} \in A_m$. Soit aussi E_m un sous-espace de dimension m de E tel que $A_m = \bar{z} + E_m$. Dans ce cas, les paramètres à estimer sont E_q et $\{z_q; q : 1, 2, \dots, k\}$.

Avec le modèle défini de la façon précédente, l'estimation par les moindres carrés s'écrit ainsi :

$$\min_{E_m, z_k} \left\{ \sum_{j=1}^k \sum_{i \in T_k} p_i \|x_i - z_j\|^2; \dim(E_m) = m, z_k - \bar{z} \in E_m \right\}$$

Comme on peut séparer la double sommation comme suit :

$$\sum_{j=1}^k \sum_{i \in T_k} p_i \|x_i - z_j\|^2 = \sum_{j=1}^k \sum_{i \in T_k} p_i \|x_i - g_j\|^2 + \sum_{j=1}^k \bar{p}_j \|g_j - z_k\|^2$$

cela revient à minimiser

$$\min_{E_m, z_k} \left\{ \sum_{j=1}^k \bar{p}_j \|g_j - z_k\|^2; \dim(E_m) = m, z_k - \bar{z} \in E_m \right\}$$

La covariance $\sigma^2 W$ du modèle étant inconnue, il faut l'estimer. Ce modèle stipule que l'ensemble des observations d'une même classe T_j suit une loi statistique non déterminée de moyenne z_j et de variance $w_{i,j}$; $\forall i \in T_j$. Dans ce cas, la matrice de covariances intra-classe ou matrice des covariances résiduelles empiriques S_r fournit donc une estimation optimale de la métrique de référence M :

$$M = \hat{W}^{-1} = S_r^{-1}$$

On en conclut, [Baccini 2007], que les paramètres E_m et z_q de ce modèle sont obtenus par l'analyse en composantes principales de (G, S_r^{-1}, \bar{D}) .

1.3.5 Métrique de Mahalanobis

L'analyse discriminante est souvent introduite dans la littérature comme un cas particulier de l'analyse canonique entre un ensemble de p variables quantitatives et un ensemble de k variables indicatrices des modalités de T . Cette relation est établie entre les deux approches car l'analyse en composantes principales basée sur le triplet (G, S_r^{-1}, \bar{D}) conduit aux mêmes vecteurs principaux que l'ACP du triplet (G, S^{-1}, \bar{D}) . Cette dernière est l'ACP des centres des classes lorsque l'espace des observations est muni de la métrique de Mahalanobis $M = S^{-1}$ et l'espace des variables de la métrique des poids des classes \bar{D} .

1.3.6 Recherche des axes factoriels discriminants

Pour effectuer l'AFD, il faut trouver le vecteur unitaire $\mu \in \mathfrak{R}^p$, tel que les projections orthogonales, g_j^q , des centres de gravité, g_j , sur le support de μ maximisent la covariance interclasse et qu'il minimise la variance intra-classes.

Pour ce faire, nous avons besoin de la variance des projections des observations, noté $\text{var}(\mu)$, sur le prolongement du vecteur unitaire μ qui est défini comme :

$$\text{var}(\mu) = \frac{1}{n} \sum_{i=1}^n [\mu(i) - \bar{\mu}]^2$$

où $\mu(i)$ est la valeur de l'abscisse de la projection orthogonale de l'observation x_i et $\bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i^T \mu$. Cela nous permet d'écrire $\text{var}(\mu)$ comme :

$$\text{var}(\mu) = \mu^T S \mu$$

En utilisant la relation qui affirme que $S = S_e + S_r$, on obtient l'équation suivante :

$$\mu^T S \mu = \mu^T S_e \mu + \mu^T S_r \mu$$

Donc, trouver le premier axe factoriel revient à maximiser, sous la contrainte $\|\mu\| = 1$, le rapport de la variance interclasse sur la variance intra-classe qui est calculée en fonction de μ . Nous pouvons aussi utiliser une des équivalences qui suivent :

$$\max_{\mu \|\mu\|=1} \frac{\mu^T S_e \mu}{\mu^T S_r \mu} = \min_{\mu \|\mu\|=1} \frac{\mu^T S_r \mu}{\mu^T S_e \mu} = \min_{\mu \|\mu\|=1} \frac{\mu^T S \mu}{\mu^T S_e \mu} = \max_{\mu \|\mu\|=1} \frac{\mu^T S_e \mu}{\mu^T S \mu}$$

[Crucianu *et al.* 2004] montrent que la solution est donnée par :

$$S_e \mu = \lambda S \mu$$

$$S^{-1} S_e \mu = \lambda \mu$$

Par les relations d'équivalences énoncées précédemment on peut aussi écrire :

$$S_e \mu = \lambda S_r \mu$$

$$S_e S_r^{-1} \mu = \lambda \mu$$

que nous pouvons réécrire comme :

$$\mu^T S_e \mu = \lambda \mu^T S_r \mu$$

$$\lambda = \frac{\mu^T S_e \mu}{\mu^T S_r \mu}$$

Comme c'est ce que nous voulons maximiser, le premier axe factoriel discriminant est le vecteur propre μ_1 correspondant à la plus grande valeur propre, λ_1 , qui appartient à la matrice S_r^{-1} .

On peut maintenant en déduire que le $\alpha^{\text{ième}}$ axe factoriel discriminant sera le vecteur propre μ_α correspondant à la valeur propre λ_α de la suite ordonnée $1 \geq \lambda_1 > \lambda_2 > \dots > \lambda_\alpha > \dots > \lambda_{k-1} > 0$, car $\mu^T S \mu = \mu^T S_e \mu + \mu^T S_r \mu$, de la matrice $S_e S_r^{-1}$. Il y aura au plus $k - 1$ valeurs propres non nulles car la matrice S_r est de rang $k - 1$.

1.3.7 Représentation des observations

Une représentation simultanée des observations et des centres de gravités des différentes classes par rapport aux mêmes axes discriminants est obtenue au moyen de la matrice C des coordonnées des observations et de la matrice \bar{C} pour les centres de gravités.

$$C = X S_r^{-1} U$$

$$\bar{C} = G S_r^{-1} U$$

où $U = [\mu_1, \mu_2, \dots, \mu_{k-1}]$ représente la matrice des vecteurs propres orthonormés de S_r^{-1} .

Les observations initiales sont projetées comme des observations supplémentaires dans le système des axes discriminants. Comme dans l'analyse en composantes principales, on peut calculer des cosinus carrés pour préciser la qualité de la représentation de chaque observation. Il est utile de différencier graphiquement la classe de chaque observation afin de pouvoir apprécier visuellement la qualité de la discrimination.

1.3.8 Représentation des variables

Chaque variable est représentée par un vecteur dont les coordonnées dans le nouveau système d'axes factoriels est une ligne de la matrice $UA^{1/2}$.

Dans ce cas, A est la matrice diagonale des valeurs propres ordonnées de S_r^{-1} .

$$A = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{k-1} \end{bmatrix}$$

La représentation des variables est utilisée pour interpréter les axes en fonction des variables initiales conjointement avec la matrice des corrélations expliquées $\Sigma_e^{-1}UA^{-1/2}$. La matrice Σ_e^{-1} est la matrice diagonale des écarts-types expliqués σ_e^j qui sont eux-mêmes les racines carrées des éléments diagonaux de la matrice S_e .

Le point essentiel est de savoir si la représentation des observations permet de faire une bonne discrimination entre les classes définies par la variable T . Si cela n'est pas le cas, l'AFD ne sert strictement à rien car les variables disponibles n'expliquent pas la variable T . Dans le cas où l'AFD permet d'obtenir

des résultats probants, le graphique des observations permet d'interpréter la discrimination en fonction des axes et le graphique des variables permet d'interpréter les axes en fonction des variables initiales.

1.3.9 Affectation d'une nouvelle observation à une classe k

Cette étape, est l'aspect décisionnel de l'AFD. Supposons une nouvelle observation x_{n+1} , située dans le même système de coordonnées que les éléments de N , pour laquelle on possède l'ensemble des mêmes variables p mesurées sur les n premières observations à l'exception de la variable T représentant sa classe d'appartenance. Nous affecterons x_{n+1} à la classe k pour laquelle la distance entre x_{n+1} et g^k est la plus petite. Cela revient à dire qu'il faut classer x_{n+1} dans le groupe k qui minimise :

$$d_{S_r^{-1}}^2(x_{n+1}, g_j), j = 1, \dots, k$$

Cette distance peut s'écrire comme :

$$d_{S_r^{-1}}^2(x_{n+1}, g_j) = \|x_{n+1} - g_j\|^2 = (x_{n+1} - g_j)^T S_r^{-1} (x_{n+1} - g_j)$$

Le problème revient à déterminer la classe d'appartenance qui maximise l'expression suivante :

$$\max_{j: 1, 2, \dots, k} \left[g_j^T S_r^{-1} x_{n+1} - \frac{1}{2} g_j^T S_r^{-1} g_j \right]$$

1.3.10 Cas de 2 groupes

Dans ce cas particulier, le nombre de facteurs discriminants à déterminer est égal à 1. Donc, il n'y a qu'une seule valeur propre non nulle λ_1 , un seul vecteur discriminant v_1 et un seul axe discriminant Δ_1 . Les deux centres de gravités g_1 et g_2 sont sur Δ_1 , de sorte que v_1 est colinéaire à $g_1 - g_2$.

L'application de la règle de décision permet d'affecter x_{n+1} au premier groupe si :

$$g_1^T S_r^{-1} x_{n+1} - \frac{1}{2} g_1^T S_r^{-1} g_1 > g_2^T S_r^{-1} x_{n+1} - \frac{1}{2} g_2^T S_r^{-1} g_2$$

De façon équivalente, x_{n+1} appartient au premier groupe si :

$$(g_1 - g_2)^T S_r^{-1} x_{n+1} > (g_1 - g_2)^T S_r^{-1} \frac{(g_1 + g_2)}{2}$$

On retrouve ainsi la formulation due à [Fisher 1936] dans les premiers travaux sur l'analyse discriminante. Si la norme de la projection de x_{n+1} sur Δ_1 est plus grande que $\frac{g_1 + g_2}{2}$ on choisit de classer la nouvelle observation dans le groupe 1; sinon, on le classe dans le groupe 2.

Il est important de remarquer que la règle de décision liée à l'AFD est simple mais elle est limitée et insuffisante dans bien des cas, notamment si les variances des classes ne sont pas identiques. De plus, elle ne tient pas compte de l'échantillonnage pour x_{n+1} ; c'est-à-dire que tous les groupes n'ont pas nécessairement la même probabilité d'occurrence. Pour ces raisons, il existe plusieurs méthodes pour construire une règle de décision qui peut être plus convenable [Baccini *et al.* 1999].

1.3.11 Évaluations des règles de décision

Dans cette section, on décrit trois méthodes courantes pour l'évaluation d'une règle de décision en analyse discriminante. Cette étape est importante pour s'assurer de la fiabilité des résultats.

La première méthode est celle de la réaffectation. Elle consiste à appliquer la règle de décision sur un échantillon d'apprentissage puis, on calcule ensuite le pourcentage de mal classées. Il est nommé le pourcentage apparent d'erreur et est fourni automatiquement par la majorité des logiciels statistiques qui sont capables d'effectuer l'analyse discriminante. Il n'est pas coûteux en temps de calcul mais il sous-estime souvent le taux d'erreur. Il est en effet biaisé car il est estimé sur les données qui ont servi à définir la règle de décision. Il est d'autant plus faible que le modèle est complexe et que la taille de l'échantillon est faible.

La deuxième méthode est celle de l'échantillon test. Elle consiste à partager l'échantillon en deux parties : la première, environ 80%, qui sert d'échantillon d'apprentissage pour produire la règle de décision et la seconde qui sert à la tester. Cette estimation de l'erreur est plus fiable et non biaisée mais nécessite un échantillon plus important. En effectuant plusieurs tirages aléatoires d'échantillons d'apprentissage et d'échantillons tests, on améliore encore l'estimation du taux d'erreur en calculant la moyenne des valeurs obtenues pour chacun des tirages.

La troisième méthode est la validation croisée. Dans cette méthode, on considère n échantillons d'apprentissage constitués en éliminant la i ème observation et cela $\forall i = 1, \dots, n$. La règle de décision qui en découle est utilisée pour affecter cette i ème observation, soit celle qui n'a pas été utilisée pour calculer la règle de décision, à l'un des différents groupes observés *a priori*. Le taux d'erreur est estimé en divisant le nombre de mal classés par n . Donc, si $n = 10$, on disposera de 10 échantillons différents de neuf observations pour

lesquels on testera la règle de décision avec l'observation qui n'a pas été utilisée dans son calcul.

1.3.12 Exemple

Toujours dans le but de faciliter la compréhension de l'analyse discriminante, cette section présente un exemple classique de la méthode appliquée sur des données relative à la qualité du vin qui a été présenté entre autre dans [Tenenhaus 1996].

Dans cet exemple, on cherche à expliquer la qualité des vins de Bordeaux par des variables météorologiques. La variable à expliquer, nommée « qualité », prend 3 modalités :

- 1= bon
- 2= moyen
- 3= médiocre

Les variables explicatives de la qualité sont :

- STEMP : Somme des températures moyennes journalières (°C)
- INSOL : Durée d'insolation (h)
- CHALEUR : Nombre de jours de grande chaleur
- HPLUIES : Hauteur des pluies (mm)

Le tableau 8 présente les données qui sont utilisées dans l'exemple. On dispose de 34 années d'observations, de 1924 à 1957, pour lesquelles aucunes des valeurs ne sont manquantes.

Le tableau 9 présente la moyenne et l'écart-type pour chacun des trois groupes et cela pour les quatre variables explicatives.

Tableau 8: Présentations des données

ANNEE	STEMP	INSOL	CHALEUR	HPLUIES	QUALITE
1924	3064	1201	10	361	2
1925	3000	1053	11	338	3
1926	3155	1133	19	393	2
1927	3085	970	4	467	3
1928	3245	1258	36	294	1
1929	3267	1386	35	225	1
1930	3080	966	13	417	3
1931	2974	1189	12	488	3
1932	3038	1103	14	677	3
1933	3318	1310	29	427	2
1934	3317	1362	25	326	1
1935	3182	1171	28	326	3
1936	2998	1102	9	349	3
1937	3221	1424	21	382	1
1938	3019	1230	16	275	2
1939	3022	1285	9	303	2
1940	3094	1329	11	339	2
1941	3009	1210	15	536	3
1942	3227	1331	21	414	2
1943	3308	1366	24	282	1
1944	3212	1289	17	302	2
1945	3361	1444	25	253	1
1946	3061	1175	12	261	2
1947	3478	1317	42	259	1
1948	3126	1248	11	315	2
1949	3458	1508	43	286	1
1950	3252	1361	26	346	2
1951	3052	1186	14	443	3
1952	3270	1399	24	306	1
1953	3198	1259	20	367	1
1954	2904	1164	6	311	3
1955	3247	1277	19	375	1
1956	3083	1195	5	441	3
1957	3043	1208	14	371	3

Tableau 9: Statistiques descriptives pour les trois groupes

qualite	1	2	3	TOTAL
COMPTAGES	11	11	12	34
MOYENNES				
stemp	3306,36	3140,91	3037,33	3157,88
insol	1363,64	1262,91	1126,42	1247,32
chaleur	28,5455	16,4545	12,0833	18,8235
hpluies	305,0	339,636	430,333	360,441
ECARTS-TYPES				
stemp	92,0568	100,045	69,3389	141,184
insol	80,306	71,9409	88,3932	126,623
chaleur	8,80186	6,7285	6,30236	10,0166
hpluies	52,2934	54,9859	104,846	91,4016

On peut ensuite déterminer les valeurs propres qui sont reliées à chacune des fonctions discriminantes. Elles sont présentées dans le tableau 10.

Tableau 10: Valeurs propres des fonctions discriminantes

Fonction discriminante	Valeur propre	Pourcentage relatif	Corrélation canonique	
1	3,27886	95,95	0,87538	
2	0,138574	4,05	0,34887	
Fonctions dérivées	Lambda de Wilks	Chi-carré	DDL	Proba.
1	0,205263	46,7122	8	0,0000
2	0,878292	3,8284	3	0,2806

Il est aussi intéressant, pour comprendre l'importance de chacune des variables dans la discrimination, de regarder les coefficients des fonctions de discrimination pour chacune des variables. Ils sont présentés dans le tableau 11. La première fonction discriminante est la suivante :

$$z_1 = 0.75 \text{ stemp} + 0.55 \text{ insol} - 0.20 \text{ chaleur} - 0.45 \text{ hpluies}$$

Tableau 11: Coefficients des fonctions de discrimination

Coefficients standardisés		
	1	2
stemp	0,750126	-0,00405015
insol	0,547064	-0,430399
chaleur	-0,198237	0,935229
hpluies	-0,445097	0,468536

On peut maintenant représenter chacune des observations dans le plan formé des deux fonctions discriminantes. C'est ce que représente la figure 9.

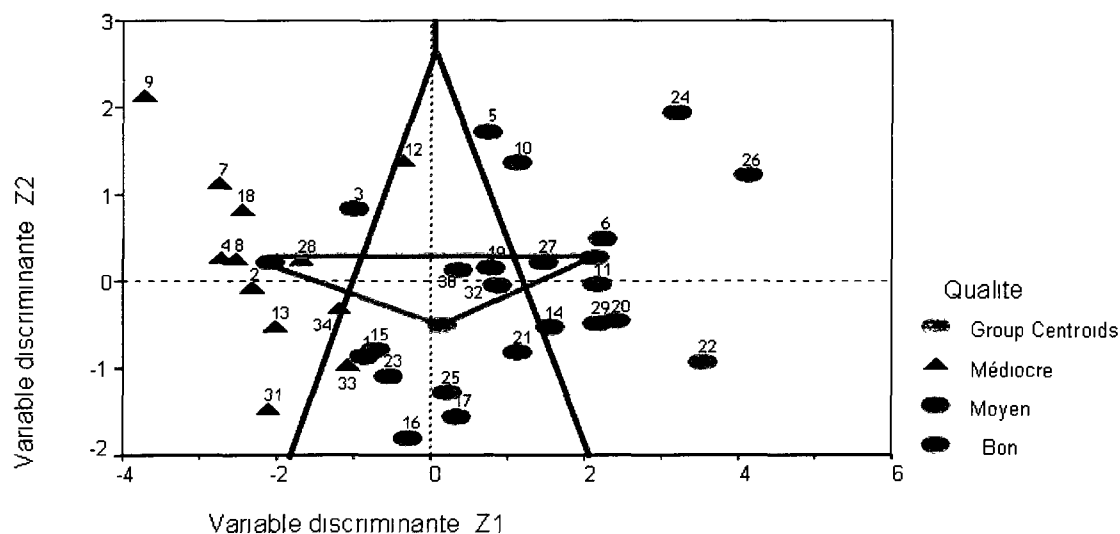


Figure 9: Représentation des observations dans le plan discriminant

De plus, on y voit les centres de gravités pour chacun des trois groupes qui sont reliés entre eux. Les médiatrices de ces trois segments de droites sont les droites frontières entre les zones d'affectation d'une nouvelle observation.

Un fois ces droites calculées, on peut regarder le pourcentage d'observations qui ont été bien classées. Ces pourcentages sont disponibles dans le tableau 12.

Tableau 12: Tableau de classification des observations

Observé qualite	Groupe Taille	Prévu 1	qualite 2	3
1	11	9 (81,82%)	2 (18,18%)	0 (0,00%)
2	11	2 (18,18%)	8 (72,73%)	1 (9,09%)
3	12	0 (0,00%)	2 (16,67%)	10 (83,33%)

Pourcentage d'observations bien classées: 79,41%				

On peut voir que 80% des observations ont été classées correctement et que deux observations ont été mal classées dans les groupes 1 et 3 et que trois observations sont dans la même situation pour le groupe 2.

Cet exemple conclut la présentation de la méthodologie dans le cadre classique. On peut maintenant passer au point de vue fonctionnel avec quelques rappels et la présentation théorique.

Chapitre 2 : Outils pour l'analyse de données fonctionnelles

Dans ce mémoire, l'intérêt est porté sur les deux outils statistiques présentés au chapitre précédent mais sous un angle particulier. En effet, dans la plupart des problèmes généralement présentés dans le domaine de la statistique, X est un vecteur aléatoire à valeurs dans un espace à dimension finie comme par exemple \mathbb{R}^n . Ici, on étudie le cas où X appartient à un espace de Hilbert, H , mais de dimension quelconque (éventuellement infinie). L'intérêt de ce type de données est qu'elles apparaissent de plus en plus fréquemment dans les problèmes concrets puisque les appareils de mesures modernes peuvent enregistrer des données qui se présentent sous la forme de courbes et qui peuvent donc être considérées comme des fonctions discrétisées en certains points. Ces données sont appelées données fonctionnelles. La figure 10 présente quatre exemples.

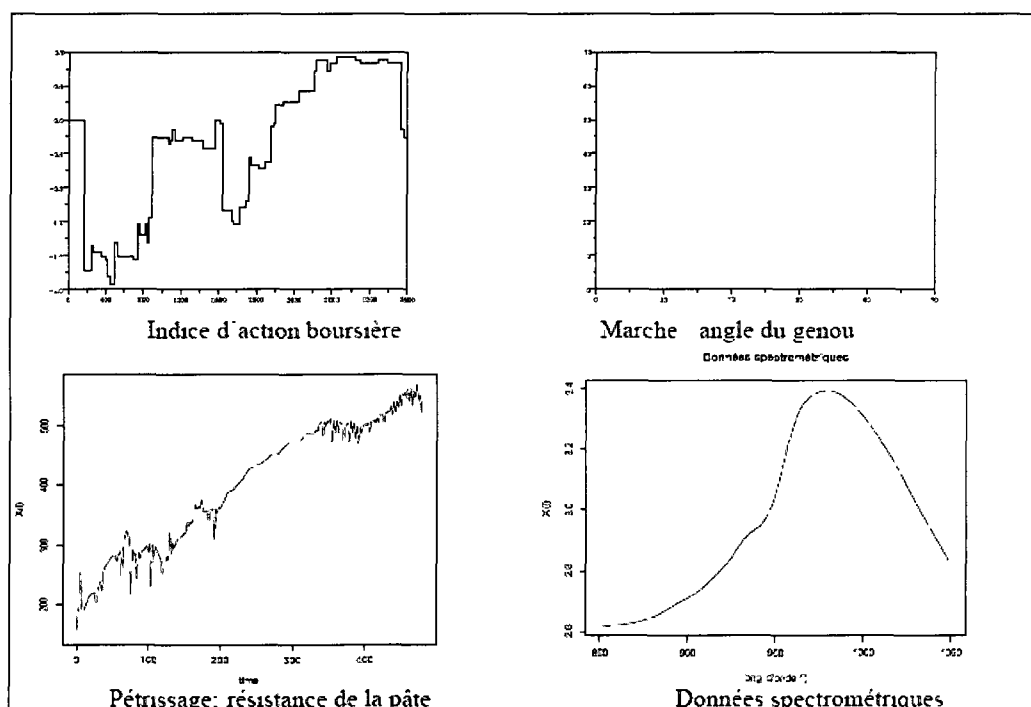


Figure 10 : Exemples de données fonctionnelles

On remarque que, même si la forme des courbes varie, on peut considérer les données comme des fonctions. Que ce soit pour l'indice d'une action boursière, pour le niveau de résistance de la pâte à pain face au pétrissage automatisé, pour l'angle du genou lors d'un pas complet ou pour le niveau d'absorbance d'un liquide face à différentes longueurs d'ondes (données spectrométriques).

L'analyse statistique des données fonctionnelles conduit à des problèmes nouveaux qui sont liés au fait que la dimension de l'espace H est infinie. Cela implique en particulier que certains problèmes qui sont résolus simplement lorsque $H = \mathbb{R}^n$ deviennent mal posés dans un espace de dimension infinie et donnent lieu, d'un point de vue pratique, à des solutions inappropriées si le caractère fonctionnel des données n'est pas pris en compte dans leur traitement.

Malgré les difficultés, la plupart des modèles statistiques classiques ont été étendus au cas fonctionnel moyennant quelques adaptations. Tout d'abord, [Deville, 1974], [Dauxois *et al.*, 1976], [Besse *et al.*, 1986] et [Besse, 1991] proposent diverses approches pour l'extension des analyses factorielles au cadre fonctionnel. Parallèlement, [Saporta, 1981] présente une étude synthétique des méthodes exploratoires d'analyse des processus. Plus tard, [Ramsay *et al.*, 1997] donnent une vision globale du traitement des données fonctionnelles avec des méthodes de régression, de discrimination et également des analyses factorielles. [Aguilera *et al.*, 1997] ont également développé un modèle linéaire de prédiction de séries chronologiques basé sur l'analyse en composantes principales d'un processus et qui utilise une approximation spline des facteurs principaux. Par ailleurs, [Cardot *et al.*, 1999] ont généralisé la régression linéaire au cadre fonctionnel alors que [Ferraty *et al.*, 2002] ont proposé une approche non paramétrique du problème de la régression fonctionnelle. Une alternative à ces approches est proposée par [Preda *et al.*, 2002] qui développent un modèle fonctionnel de régression PLS (Partial Least Squares). Il s'agit d'une méthode itérative de régression linéaire qui se révèle particulièrement efficace dans le cadre fonctionnel. Enfin, dans [Preda *et al.*, 2005], les auteurs proposent une

application de la régression PLS dans le cadre de l'analyse discriminante pour une réponse binaire. [Dauxois et al., 2001], [Ferré et Yao, 2003] et [Ferré and Yao, 2005] développent un modèle semi-paramétrique pour variable aléatoire hilbertienne qui est une version fonctionnelle de la SIR (Sliced Inverse Regression) ([Li, 1991]). Ce dernier modèle peut être utilisé aussi bien à des fins de régression que de discrimination ([Dauxois et al., 2001] et [Ferré et Villa, 2005]). Par ailleurs, [Rossi et al., 2004], [Rossi et Conan-Guez, 2005] et [Rossi et al., 2005] ont développé des méthodes de traitement de données fonctionnelles par réseaux de neurones.

Dans le domaine de la discrimination, les dernières années ont été aussi riches en travaux. Un certain nombre d'entre eux présentent des approches par pénalisation de l'opérateur de covariance. Cette méthode, qui permet une régularisation du problème initial, a été développée, à l'origine par [Ivanov, 1962], [Tihonov, 1963a] et [Tihonov, 1963b]. Elle a été appliquée avec succès à l'analyse en composantes principales ([Pezzulli et al., 1993]; [Silverman, 1996]), à l'analyse discriminante ([Hastie et al., 1994] et [Hastie et al., 1995]) et à l'analyse canonique ([Leurgans et al., 1993]). Une approche alternative aux méthodes de régularisation est le filtrage qui consiste à projeter les données sur une base de fonctions préalablement fixées. C'est ce que font, par exemple, [Biau et al., 2005] qui développent une méthode de discrimination, dans des espaces de Hilbert, par le plus proches voisins. Enfin, [James et al., 2001] montrent, dans le cadre de l'analyse discriminante, la limite de ces deux approches, particulièrement lorsque les données sont échantillonnées de manière non uniforme. Ils proposent alors un modèle dans lequel la fonction sous-jacente, et non les observations, est exprimée par filtrage sur une base spline.

Avant de présenter en détails les méthodes qui seront utilisées, on retrouve dans ce chapitre, quelques notations et éléments de base essentiels à l'analyse de

données fonctionnelles. Certains des éléments présentés sont sûrement familiers, mais il est bon, dans ce contexte, de faire un bref rappel.

2.1 Dérivées et intégrales

Soit une fonction $x(t)$ à valeurs réelles définie sur \mathfrak{R} ou un sous-ensemble de \mathfrak{R} .

La notation usuelle pour une dérivée d'ordre m d'une fonction $x(t)$ est :

$$\frac{d^m x(t)}{dt^m}$$

La dérivée est un opérateur qui agit sur la fonction $x(t)$ pour produire une nouvelle fonction notée $x'(t)$. La dérivée d'ordre 0, que l'on notera :

$$\frac{d^0 x(t)}{dt^0}$$

fait référence à la fonction $x(t)$ elle même. L'utilisation d'une dérivée d'ordre négatif peut aussi être possible, cette notation fait référence à l'intégrale indéfinie de la fonction $x(t)$. L'intégrale est la fonction inverse de la dérivée :

$$\frac{d^m}{dt^m} \left[\frac{d^{-m}}{dt^{-m}} x(t) \right] = \frac{d^0 x(t)}{dt^0} = x(t)$$

L'intégrale a aussi une notation propre :

$$\int x(t) dt$$

qui représente l'intégrale indéfinie de $x(t)$. Nous pouvons aussi spécifier les bornes d'intégration, tel que :

$$\int_a^b x(t) dt$$

ce qui signifie que l'on évaluera la valeur de l'intégrale entre le point $t = a$ et $t = b$ de la fonction $x(t)$. On appelle ce type d'intégrale, intégrale définie.

2.2 Produit scalaire et espace L^p

Avant d'introduire la notion du produit scalaire, il faut définir correctement l'espace dans lequel ce dernier est défini. Cela amène la présentation de l'espace L^2 et plus généralement les espaces L^p .

Cette section est inspirée des textes de [Brezis 2005] et de [Samuelides *et al.* 1989].

2.2.1 L'espace L^p

Soient (Ω, F, μ) un espace mesuré, $1 \leq p \leq \infty$ et $f \in M = M(\Omega, F)$ où M désigne l'ensemble des fonctions mesurables de (Ω, F) dans \mathfrak{R} (c'est-à-dire $f : \Omega \rightarrow \mathfrak{R}$, mesurable). On remarque que $|f|^p \in M_+$, $|f|^p = \varphi \circ f$ où φ est la fonction continue définie par $\varphi(s) = |s|^p$ pour tout $s \in \mathfrak{R}$. La quantité $\int |f|^p d\mu$ est donc bien définie et appartient à $\overline{\mathfrak{R}}_+$. Ceci permet de définir les espaces de fonctions de puissance $p^{\text{ème}}$ intégrable. On retrouve, pour $p = 1$, la définition de l'espace des fonctions intégrables.

Définition 2.1 : Soient (Ω, F, μ) un espace mesuré, $1 \leq p \leq \infty$ et f une fonction définie de Ω dans \mathfrak{R} , mesurable (on a donc $|f|^p \in M_+$).

1. On dit que $f \in \mathbf{L}^p = \mathbf{L}^p(\Omega, F, \mu)$ si $\int |f|^p du < \infty$. On pose alors :

$$\|f\|_p = \left(\int |f|^p du \right)^{\frac{1}{p}}$$

2. On dit que $f \notin \mathbf{L}^p = \mathbf{L}^p(\Omega, F, \mu)$ si $\int |f|^p du = \infty$ et on pose alors :

$$\|f\|_p = \infty$$

Pour la suite, on a besoin d'introduire la notion d'égalité presque partout : soient f et g deux fonctions mesurables appartenant à M . On dit que f et g sont égales presque partout, que l'on note $f \stackrel{pp}{=} g$, si :

$$\{x \in \Omega : f(x) \neq g(x)\}$$

est un ensemble négligeable c'est-à-dire, de mesure nulle. En d'autres termes :

$$\mu \{x \in \Omega : f(x) \neq g(x)\} = 0$$

Cette notion permet de définir les espaces L^p .

Définition 2.2 : Soit (Ω, F, μ) un espace mesuré, $1 \leq p \leq \infty$:

1. On définit l'espace $L^p_{\mathfrak{R}}(\Omega, F, \mu)$ comme l'ensemble des classes d'équivalence des fonctions de \mathbf{L}^p pour la relation d'équivalence égalité

presque partout ($\stackrel{pp}{=}$). En l'absence d'ambiguïté on notera L^p l'espace $L^p_{\mathfrak{R}}(\Omega, F, \mu)$.

2. Soit $G \in L^p_{\mathfrak{R}}(\Omega, F, \mu)$. On pose : $\|G\|_p = \|g\|_p$ si $g \in G = \tilde{g} = \left\{ f \in L^p; f \stackrel{pp}{=} g \right\}$

2.2.2 Produit scalaire

Dans cette partie, le produit scalaire et les éléments qui en découlent sont définis pour des fonctions qui appartiennent à L^2 . On verra, au chapitre 3, pourquoi il est essentiel que les fonctions analysées soient de carrés intégrables.

Le produit scalaire pour deux fonctions réelles, $x, y \in L^2$ est défini de la façon suivante :

$$\langle x, y \rangle = \int xy du$$

Ce produit scalaire possède les propriétés suivantes :

Symétrie : $\langle x, y \rangle = \langle y, x \rangle \forall x \text{ et } y$

Positivité : $\langle x, x \rangle \geq 0 \forall x$ et $\langle x, x \rangle = 0$ si et seulement si $x \stackrel{pp}{=} 0$

Bilinéarité : $\forall a, b \in \mathfrak{R} \quad \langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$

Une notion mathématique qui est liée au produit scalaire est la notion de la norme, définie par :

$$\|x\| = \sqrt{\int x^2 du}$$

Cette norme est appelée la norme L_2 . Elle est définie, en relation avec le produit scalaire, comme suit :

$$\|x\|^2 = \langle x, x \rangle = \int x^2 du$$

Cette relation nous permet de déduire quelques propriétés de la norme :

$$\|x\| \geq 0 \text{ et } \|x\| = 0 \text{ si et seulement si } x = 0$$

$$\|ax\| = |a|\|x\| \quad \forall a \in \mathbb{R}$$

$$\|x + y\| \leq \|x\| + \|y\|$$

Des propriétés du produit scalaire on retrouve l'inégalité de Cauchy-Schwarz :

$$|\langle x, y \rangle| \leq \|x\| \times \|y\| = \sqrt{\langle x, x \rangle \langle y, y \rangle}$$

Cela nous permet d'énoncer ce que l'on appelle l'inégalité du cosinus :

$$-1 \leq \frac{\langle x, y \rangle}{\|x\| \times \|y\|} \leq 1$$

On suppose que x et y sont des vecteurs à n dimensions dans un espace muni de la métrique euclidienne usuelle. Dans ce cas, l'angle, θ , entre les deux vecteurs peut être trouvé avec la formule suivante :

$$\cos(\theta) = \frac{\langle x, y \rangle}{\|x\| \times \|y\|}$$

On peut utiliser le même concept pour le cosinus de l'angle entre deux fonctions. Avec cette notion d'angle entre les fonctions, le cas où $\langle x, y \rangle = 0$ devient particulièrement intéressant car, on dit que x et y sont orthogonales. De plus, par la propriété de bilinéarité, on peut dire que l'orthogonalité demeure peu importe l'échelle utilisée. Cette notion est très importante pour les concepts de qualité de la représentation et de la projection discutés plus loin.

2.3 Espace vectoriel

Une autre notion importante est celle de l'espace vectoriel. Elle est définie dans cette sous-section.

Définition 2.3 : Un corps est un ensemble K muni de deux lois internes notées en général $+$ et \times vérifiant :

- $(K, +)$ forme un groupe commutatif dont l'élément neutre est noté 0
- $(K \setminus \{0\}, \times)$ forme un groupe multiplicatif.
- La multiplication est distributive pour l'addition (à gauche comme à droite)

Le corps K est noté $(K, +, \times)$

Définition 2.4 : En algèbre, un groupe abélien, ou groupe commutatif, est un ensemble non vide muni d'une loi de composition interne $(G, *)$ dont cette loi $*$ est commutative, c'est-à-dire que $\forall \alpha, \beta \in G : \alpha * \beta = \beta * \alpha$.

Définition 2.5 : On appelle loi de composition interne sur un ensemble E toute application $*$ de $E \times E \rightarrow E$. Un ensemble E muni d'une loi de composition interne $*$ constitue une structure algébrique et est notée $(E, *)$.

Définition 2.6 : Soient A et B des ensembles. On appelle loi externe sur B une application $\theta : A \times B \rightarrow B$. Par convention, nous utilisons la notation suivante : si $\alpha \in A$ et $x \in B$; $\theta(\alpha, x) = \alpha x$.

Définition 2.7 : Soit $(K, +, \times)$ un corps et soit $(E, +)$ un groupe abélien. Soit aussi $\theta : K \times E \rightarrow E$ une loi externe sur E . Le triplet $(E, +, \times)$ a une structure d'espace vectoriel sur K si :

- $\forall \alpha \in K, \forall x, y \in E : \alpha(x + y) = \alpha x + \alpha y$
- $\forall \alpha, \beta \in K, \forall x \in E : (\alpha + \beta)x = \alpha x + \beta x$
- $\forall \alpha, \beta \in K, \forall x \in E : (\alpha \times \beta)x = \alpha(\beta \times x)$

Dans cette définition, K est appelé le corps de base de l'espace vectoriel E . Bien que K puisse être n'importe quel corps, dans cette partie du mémoire K est représenté les réels, noté \mathbb{R} .

Définition 2.8: Soit K un corps et E un espace vectoriel sur K . Soit V un sous ensemble de E . V est un sous espace vectoriel de E si :

- $(V, +)$ est un sous groupe de $(E, +)$
- Si $\forall \alpha \in K, \forall x \in V : \alpha x \in V$

En vérifiant les axiomes définissant un espace vectoriel, on peut démontrer qu'un sous-espace vectoriel d'un espace vectoriel sur un corps K est aussi un espace vectoriel sur K .

Proposition 2.1 : Soient K un corps, E un espace vectoriel sur K et V un sous-ensemble de E . On peut dire que V est un sous-espace vectoriel de E si et seulement si :

$$\forall \alpha, \beta \in K, \forall x, y \in V : \alpha x + \beta y \in V$$

Définition 2.9 : Soit E un espace vectoriel sur un corps K . Soit A une partition de E . On suppose que $A = \{\lambda_i; i \in I\}$ où I est un ensemble permettant d'indexer A . On appelle combinaison linéaire des éléments de A tout élément de E pouvant s'écrire comme :

$$x = \sum_{i \in I} \lambda_i x_i$$

Définition 2.10 : Soit $A = \{\lambda_i; i \in I\}$ une partition de E un espace vectoriel sur K . A est libre dans E ou une famille libre des vecteurs de E si :

$$\sum_{i \in I} \lambda_i x_i = 0 \Rightarrow \lambda_i = 0 \forall i .$$

Dans le cas contraire la partition de A est dite liée dans E .

Définition 2.11 : Soit E un espace vectoriel sur K . Soit A un sous-ensemble de E . L'ensemble des combinaisons linéaires des éléments de A est un sous-espace vectoriel de E appelé sous-espace vectoriel de E engendré par A .

Définition 2.12 : Soit E un espace vectoriel sur K . Soit V un sous-espace vectoriel de E . Soient A une famille de vecteurs de V . Cette famille est dite génératrice de V si l'espace vectoriel engendré par A est égal à V .

Définition 2.13 : Soit E un espace vectoriel sur K . Soient I un ensemble et $A = \{\lambda_i; i \in I\}$ une famille d'éléments de E . Cette famille est une base de E si elle est à la fois libre dans E et génératrice de E .

Proposition 2.2 : Soit E un espace vectoriel sur K . Soit $(e_i)_{i \in I}$ une base de E . Pour un élément de E il existe une unique famille $\{\lambda_i; i \in I\}$ de scalaire de I tel que :

$$x = \sum_{i \in I} \lambda_i e_i$$

Dans ce cas, λ_i s'appelle la $i^{\text{ème}}$ coordonnée du vecteur x relativement à la base $(e_i)_{i \in I}$.

Définition 2.14 : Soit K un corps et E un espace vectoriel sur K . E est un espace vectoriel de dimension finie s'il possède une famille génératrice de cardinal fini. Dans le cas contraire, E est dit de dimension infinie.

Proposition 2.3 : Soit E un espace vectoriel sur un corps K . Si E est de dimension finie, alors E possède une base.

Définition 2.15 : Soit E un espace vectoriel de dimension finie sur un corps K . Si E est réduit à son élément nul, alors on dit que la dimension de E est 0. Sinon, on appelle dimension de E le cardinal d'une base de E que l'on note $\dim(E)$.

2.3.1 Espace vectoriel de dimension infinie

Pour la suite de ce mémoire, on aura aussi besoin de quelques résultats sur les espaces vectoriels de dimension infinie.

Proposition 2.4 : Soit K un corps et E un espace vectoriel sur K . Soit V un sous-espace vectoriel de E . Soient aussi W_1 un sous-espace vectoriel de E tel que

$V \cap W_1 = \{0\}$ et W_2 un sous-espace vectoriel de E tel que $V + W_2 = E$. Alors il existe un supplémentaire W de V contenu dans W_2 et contenant W_1 .

De cette dernière proposition, on peut en déduire le corollaire suivant :

Soit K un corps et soit E un espace vectoriel sur K . Soit aussi V un sous espace vectoriel de E . Alors V possède un supplémentaire dans E .

Proposition 2.5 : Soit E un espace vectoriel non réduit à l'ensemble vide sur le corps K . E possède une base.

Théorème de la base incomplète en dimension infinie : Soit $(e_i)_{i \in I}$ une partie génératrice de E . Soit I' un sous ensemble de I tel que $(e_i)_{i \in I'}$ est libre dans E . Alors il existe $I' \subset I'' \subset I$ tel que $(e_i)_{i \in I''}$ soit une base de E .

2.4 Statistiques descriptives

Une analogie du produit scalaire réside dans l'expression des statistiques descriptives couramment utilisées : moyenne, variance, covariance et corrélation.

Soient x et y deux fonctions de L^2 , définie sur l'intervalle $[0, T] \in \Omega \in \mathfrak{R}$. Le produit scalaire entre ces deux fonctions est alors donné par :

$$\langle x, y \rangle = \int_0^T xy du$$

La moyenne d'une fonction x , noté \bar{x} , est définie de la façon suivante :

$$\bar{x} = \frac{\langle x, 1 \rangle}{\|1\|^2}$$

$$\bar{x} = \frac{\int_0^T x \times 1 du}{\int_0^T 1 du}$$

$$\bar{x} = \frac{\int_0^T x du}{T}$$

En utilisant les mêmes notations, on peut exprimer la variance de la fonction x , notée s_x^2 par :

$$s_x^2 = \frac{\|x - \bar{x}1\|^2}{\|1\|^2}$$

La covariance entre les fonctions x et y , notée $s_{x,y}$, est exprimé de la façon suivante :

$$s_{x,y} = \frac{\langle x - \bar{x}1, y - \bar{y}1 \rangle}{\|1\|^2}$$

Il reste à définir la corrélation entre les fonctions x et y . Cette valeur est notée $r_{x,y}$ et a pour expression:

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

$$r_{x,y} = \frac{\langle x - \bar{x}1, y - \bar{y}1 \rangle}{\|1\|^2} = \frac{\langle x - \bar{x}1, y - \bar{y}1 \rangle}{\sqrt{\frac{\|x - \bar{x}\|^2}{\|1\|^2}} \times \sqrt{\frac{\|y - \bar{y}\|^2}{\|1\|^2}}}$$

$$r_{x,y} = \frac{\langle x - \bar{x}1, y - \bar{y}1 \rangle}{\sqrt{\|x - \bar{x}\|} \times \sqrt{\|y - \bar{y}\|}}$$

2.5 Espace de Hilbert

Avant de décrire les espaces de Hilbert, quelques notions essentielles à la compréhension sont nécessaires et se retrouvent dans la sous-section suivante. De plus, bien que plusieurs des résultats présentés sont aussi vrais pour des éléments du corps des complexes, on ne s'intéresse dans ce texte qu'aux cas où les éléments sont des réels.

2.5.1 Définitions

Définition 2.16 : une suite de réels (r_n) est dite de Cauchy lorsque les termes de la suite sont tels que $\forall \varepsilon > 0; \exists n \in \mathbb{N}; \forall p, q > n$ on ait:

$$|r_p - r_q| \leq \varepsilon$$

où encore :

$$\lim_{n \rightarrow \infty} \sup_{p, q > n} |r_p - r_q| = 0$$

Définition 2.17 : un espace métrique M est dit complet ou espace complet si toute suite de Cauchy de M a une limite dans M, c'est-à-dire converge dans M.

La propriété de complétude dépend de la distance. Il est donc important de toujours préciser la distance que l'on utilise quand on parle d'espace complet.

Définition 2.18 : un espace vectoriel normé $(H, \|\cdot\|)$ sur les réels ou sur les complexes, est un espace de Hilbert si sa norme est déduite d'un produit scalaire et s'il est complet.

Définition 2.19 : Une partie G de H est dite dense dans H si :

$$\forall h \in H ; \forall \varepsilon > 0 ; \exists g \in G ; \|g - h\| < \varepsilon$$

en d'autres termes si $\overline{G} = H$ où \overline{G} est l'adhérence de G

Définition 2.20 : Une partie F de H est dite totale si l'ensemble des combinaisons linéaires finies des éléments de F est dense dans H .

Définition 2.21 : Une famille $\{e_i, i \in I\}$ d'éléments de H est dite orthonormée si

$$\forall i, j \in I ; \langle e_i, e_j \rangle = d_{ij}$$

où d_{ij} est égal à 1 si $i=j$ et 0 sinon.

Définition 2.22 : Une base hilbertienne de H ou base orthonormée est une famille orthonormée et totale dans H .

2.5.2 Séparabilité et bases hilbertiennes

Rappelons qu'un espace métrique est séparable s'il possède un sous-ensemble dénombrable dense dans M . La plupart des espaces de Hilbert utilisés en pratique sont séparables. Tout espace de Hilbert possédant une suite totale

$(f_k)_{k \in \mathbb{N}}$ est séparable. Il suffit pour le voir de prendre comme partie dénombrable dense l'ensemble des combinaisons linéaires finies des f_k à coefficients rationnels.

Proposition 2.6 : Tout espace de Hilbert séparable de dimension infinie possède une base hilbertienne dénombrable.

Démonstration : Soit $\{g_n\}_{n \in \mathbb{N}}$ une suite dense dans H . En utilisant le théorème de Bolzano-Weierstrass [Wagschal 1995] qui dit que de toute suite réelle bornée, on peut extraire une sous-suite convergente, on peut en extraire par récurrence une sous-suite $(g_{n_i})_{i \in \mathbb{N}}$ telle que :

- 1) Pour tout $i \in \mathbb{N}$, $(g_{n_j}; 0 \leq j \leq i)_{j \in \mathbb{N}}$ est une famille libre.
- 2) Pour tout $i \in \mathbb{N}$, $\text{Vect}(g_k; 0 \leq k \leq n_i) = \text{Vect}(g_{n_j}; 0 \leq j \leq i)_{j \in \mathbb{N}} = E_i$.
- 3) $E = \bigcup_{i \in \mathbb{N}} E_i \supset \{g_n\}_{n \in \mathbb{N}}$.

où $\text{Vect}(g_k; 0 \leq k \leq n_i)$ représente le sous-espace vectoriel engendré par la famille est l'ensemble de toutes les combinaisons linéaires finies de la famille $(g_k; 0 \leq k \leq n_i)$.

En orthogonalisant la suite $(g_{n_i})_{i \in \mathbb{N}}$ par le procédé de Gram Schmidt on construit une suite orthonormée $(f_{n_i})_{i \in \mathbb{N}}$ telle que pour tout $i \in \mathbb{N}$, $(f_{n_j})_{0 \leq j \leq i}$ soit une base orthonormée de l'espace de dimension finie E_i . Pour montrer que la suite $(f_{n_i})_{i \in \mathbb{N}}$ est une base hilbertienne de H , il reste à vérifier qu'elle est totale dans H .

Soit f quelconque dans H et soit $\varepsilon > 0$. En vertu de la propriété 3 et de la densité de $\{g_n\}_{n \in \mathbb{N}}$ dans H , il existe un élément $g \in E$ tel que $\|f - g\| < \varepsilon$. Comme E est

l'union des E_i , il existe un indice $i \in \mathbb{N}$ tel que $g \in E_i$. Or les $(f_{n_j})_{0 \leq j \leq i}$ forment une base de l'espace de dimension finie E_i donc g se décompose dans cette base sous la forme :

$$g = \sum_{j=0}^i a_j f_{n_j}.$$

Bien sûr, g , i et les a_j dépendent de f et de ε , mais le raisonnement est valable pour tout f et tout ε . La suite $(f_{n_i})_{i \in \mathbb{N}}$ est donc totale dans H .

Proposition 2.7 : Si H est un espace de Hilbert séparable, toute base orthonormée de H est au plus dénombrable.

Démonstration : Soit $(e_i)_{i \in I}$ une base orthonormée dans H . Pour prouver que I est au plus dénombrable, il suffit de montrer l'existence d'une injection de I dans \mathbb{N} . Remarquons d'abord que si i et j sont deux éléments distincts de I :

$$\|e_i - e_j\| = \varepsilon = \sqrt{2}$$

Par la séparabilité de H il existe un ensemble dénombrable $\{g_n\}_{n \in \mathbb{N}}$ dense dans H . Pour chaque $i \in I$, il existe alors un entier n_i tel que :

$$\|e_i - g_{n_i}\| < \frac{\sqrt{2}}{3}$$

En choisissant un entier n_i pour chaque $i \in I$ parmi tous ceux vérifiant l'équation précédente, on définit une application $\varphi : I \rightarrow \mathbb{N}, i \mapsto n_i$.

En vertu de l'inégalité du triangle, on a alors pour $i \neq j$:

$$\sqrt{2} = \|e_i - e_j\| \leq \|e_i - g_{n_i}\| + \|g_{n_i} - g_{n_j}\| + \|g_{n_j} - e_j\| \leq \|g_{n_i} - g_{n_j}\| + \frac{2\sqrt{2}}{3}$$

Cela revient à dire que :

$$\|g_{n_i} - g_{n_j}\| \geq \frac{\sqrt{2}}{3}$$

On en déduit que $g_{n_i} \neq g_{n_j}$ et donc que $n_i \neq n_j$. Ainsi pour tous i, j distincts dans I , $\varphi(i) \neq \varphi(j)$ ce qui exprime l'injectivité de φ .

Il est important de noter que cette même démonstration prouve que si $\{e_i\}_{i \in \mathbb{N}}$ est seulement une famille orthonormée dans H séparable, alors I est au plus dénombrable.

Corollaire 2.1 : Soit H un espace de Hilbert séparable et $\{e_i\}_{i \in \mathbb{N}}$ une base hilbertienne. Alors ou bien H est de dimension finie et $\text{card } I = \dim H$, ou bien H est de dimension infinie et I est exactement dénombrable (i.e. en bijection avec \mathbb{N}).

Théorème 2.1 : Soit $\{e_i\}_{i \in \mathbb{N}}$ une famille orthonormée de H . Alors pour tout $f \in H$, toute partie finie J non nulle de I et tous scalaires a_j , $j \in J$

$$\left\| f - \sum_{j \in J} c_j(f) e_j \right\| \leq \left\| f - \sum_{j \in J} a_j e_j \right\|$$

avec :

$$c_j(f) = \langle f, e_j \rangle$$

L'égalité n'ayant lieu que si et seulement si $a_j = c_j(f)$ pour tout $j \in J$.

Démonstration : En réécrivant $\left\| f - \sum_{j \in J} a_j e_j \right\|$ comme :

$$\left\| f - \sum_{j \in J} a_j e_j \right\| = \left\| f - \sum_{j \in J} (a_j - c_j(f) + c_j(f)) e_j \right\|$$

et en appliquant la formule suivante :

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$$

on obtient :

$$\begin{aligned} \left\| f - \sum_{j \in J} a_j e_j \right\|^2 &= \left\| \left(f - \sum_{j \in J} c_j(f) e_j \right) + \left(\sum_{j \in J} (c_j(f) - a_j) e_j \right) \right\|^2 = \\ &= \left\| f - \sum_{j \in J} c_j(f) e_j \right\|^2 + \left\| \sum_{j \in J} (c_j(f) - a_j) e_j \right\|^2 + 2 \left\langle f - \sum_{j \in J} c_j(f) e_j, \sum_{j \in J} (c_j(f) - a_j) e_j \right\rangle \end{aligned}$$

Le produit croisé dans cette expression est nul car en développant le produit scalaire par rapport à son deuxième argument, on obtient :

$$\sum_{k \in J} (c_k(f) - a_k) \left\langle f - \sum_{j \in J} c_j(f) e_j, e_k \right\rangle = \sum_{k \in J} (c_k(f) - a_k) (\langle f, e_k \rangle - c_k(f)) = 0$$

D'autre part, grâce à l'orthonormalité des e_j on a :

$$\left\| \sum_{j \in J} (c_j(f) - a_j) e_j \right\|^2 = \sum_{j \in J} |c_j(f) - a_j|^2$$

Finalement :

$$\left\| f - \sum_{j \in J} a_j e_j \right\|^2 = \left\| f - \sum_{j \in J} c_j(f) e_j \right\|^2 + \sum_{j \in J} |c_j(f) - a_j|^2$$

$$\left\| f - \sum_{j \in J} a_j e_j \right\|^2 \geq \left\| f - \sum_{j \in J} c_j(f) e_j \right\|^2$$

l'égalité n'étant possible si et seulement si $\sum_{j \in J} |c_j(f) - a_j|^2 = 0$, autrement dit $a_j = c_j(f)$ pour tout $j \in J$.

Interprétation géométrique : le vecteur $\pi_J(f) = \sum_{j \in J} c_j(f) e_j$ est la projection orthogonale de f sur le sous espace vectoriel de dimension finie $E_J = \text{Vect}(e_j, j \in J)$. Il est orthogonal à $f - \pi_J(f)$. Définissons la distance de f à E_J par :

$$d(f, E_J) = \inf_{g \in E_J} \|f - g\|$$

Le théorème précédent nous dit alors que cet infimum est un minimum et que ce minimum est atteint en l'unique point $g = \pi_J(f)$. Ainsi la projection orthogonale de $f \in E$ sur E_J est la meilleure approximation de f par un élément de E_J . Il est clair que ce résultat se généralise immédiatement avec n'importe quel sous-espace vectoriel de dimension finie de H puisqu'un tel espace possède toujours

une base orthonormée et que seuls les éléments e_j , indexés par J ont été utilisés dans la preuve du théorème.

Corollaire 2.2 : Soit $\{e_i\}_{i \in \mathbb{N}}$ une famille orthonormée de H . Alors pour tout $f \in H$, toutes parties finies $K \subset J$ non vides de I :

$$\left\| f - \sum_{j \in J} c_j(f) e_j \right\| \leq \left\| f - \sum_{j \in K} c_j(f) e_j \right\|$$

Inégalité de Bessel : Si $\{e_i\}_{i \in \mathbb{N}}$ est une famille orthonormée au plus dénombrable de H , alors $\forall f \in H$:

$$\sum_{i \in I} |\langle f, e_i \rangle|^2 \leq \|f\|^2$$

Démonstration : Pour toute partie finie J de I :

$$\begin{aligned} \left\langle f - \sum_{j \in J} c_j(f) e_j, \sum_{j \in J} c_j(f) e_j \right\rangle &= \sum_{k \in J} c_k(f) \left\langle f - \sum_{j \in J} c_j(f) e_j, e_k \right\rangle = \\ &= \sum_{k \in J} c_k(f) (\langle f, e_k \rangle - c_k(f)) = 0 \end{aligned}$$

Cette orthogonalité permet d'écrire :

$$\|f\|^2 = \left\| f - \sum_{j \in J} c_j(f) e_j \right\|^2 + \left\| \sum_{j \in J} c_j(f) e_j \right\|^2 = \left\| f - \sum_{j \in J} c_j(f) e_j \right\|^2 + \sum_{j \in J} |c_j(f)|^2$$

d'où :

$$\sum_{j \in J} |c_j(f)|^2 \leq \|f\|^2$$

Cette inégalité étant vraie pour toute partie finie J de I , on en déduit l'inégalité de Bessel en prenant le sup sur J .

Corollaire 2.3 : Soit $\{e_k\}_{k \in \mathbb{N}}$ une suite orthonormée dans H et pour tout $f \in H$, on pose $c_k(f) = \langle f, e_k \rangle$, la coordonnée de f sur le $k^{\text{ème}}$ élément de la base. La série suivante :

$$\sum_{k=0}^{\infty} c_k(f) e_k \text{ converge commutativement dans } H.$$

La convergence commutative dans H est définie de la façon suivante :

1. $\exists g \in H$ tel que $\lim_{n \rightarrow \infty} \left\| g - \sum_{k=0}^n c_k(f) e_k \right\| = 0$
2. $\forall \tau : \mathbb{N} \rightarrow \mathbb{N}, \lim_{n \rightarrow \infty} \left\| g - \sum_{k=0}^n c_{\tau(k)}(f) e_{\tau(k)} \right\| = 0$

Identité de Parseval : Soit H un espace de Hilbert séparable et si $\{e_i\}_{i \in I}$ est une base hilbertienne de H , on a l'identité de Parseval :

$$\forall f \in H, \|f\|^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2$$

Démonstration : Si $\{e_k\}_{k \in \mathbb{N}}$ est une base hilbertienne. On a en vertu du corollaire précédent et de l'inégalité de Bessel :

$$\forall f \in H, \sum_{k=0}^n c_k(f) e_k \xrightarrow{n \rightarrow \infty} f$$

et donc, en raison de la continuité de la norme sur H , on en déduit la convergence dans \mathfrak{H}_+ :

$$\left\| \sum_{k=0}^n c_k(f) e_k \right\|^2 \xrightarrow{n \rightarrow \infty} \|f\|^2$$

par orthonormalité des e_k :

$$\left\| \sum_{k=0}^n c_k(f) e_k \right\|^2 = \sum_{k=0}^n |c_k(f)|^2$$

et en reportant e_k dans l'équation précédente on en déduit l'identité de Parseval :

$$\forall f \in H, \|f\|^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2$$

Théorème de l'unicité du développement : Soit $\{e_i\}_{i \in I}$ une base hilbertienne de l'espace de Hilbert séparable H et $\{a_i\}_{i \in I}$ une famille de scalaires telle que la série $\sum_{i \in I} a_i e_i$ converge vers un élément $f \in H$ pour au moins un ensemble de $i \in I$. Alors

$$a_i = \langle f, e_i \rangle \quad \forall i \in I$$

Démonstration : Il suffit de faire la preuve lorsque $I = N$. Par hypothèse nous avons donc :

$$f_n = \sum_{k=0}^n a_k e_k \xrightarrow{n \rightarrow \infty} f$$

D'autre part, pour tous j et n dans N , on a par orthonormalité des e_k :

$$\langle f_n, e_j \rangle = \begin{cases} 0 & \text{si } n < j \\ a_j & \text{si } n \geq j \end{cases}$$

La suite de réels $\{\langle f_n, e_j \rangle\}_{n \in N}$ est donc constante et égale à a_j à partir du rang j . Elle converge ainsi vers a_j .

D'autre part, $\langle \bullet, e_j \rangle$ est une forme linéaire continue sur H (par Cauchy Schwarz) et f_n tend vers f pour la topologie déduite de la mesure sur H . Ce qui implique que :

$$\langle f_n, e_j \rangle \xrightarrow{n \rightarrow \infty} \langle f, e_j \rangle$$

Par unicité de la limite dans les complexes ([Leborgne 1991]), on en déduit que :

$$a_i = \langle f, e_i \rangle \forall i \in I$$

Théorème de Riez-Fischer : Tout espace de Hilbert de dimension infinie et séparable est isomorphe à $l^2(N)$.

Démonstration : Rappelons que l'espace $l^2(N)$ est l'espace des suites $u = (u_k)_{k \in N}$ de nombres réels telles que :

$$\|u\|_2^2 = \sum_{k \in N} |u_k|^2 < \infty$$

et que c'est un espace de Hilbert pour le produit scalaire défini comme suit :

$$\langle u, v \rangle = \sum_{k \in N} u_k v_k$$

Soit H un espace de Hilbert de dimension infinie et séparable. Il possède alors une base hilbertienne exactement dénombrable que l'on peut toujours, quitte à la réindexer, écrire sous la forme $\{e_k\}_{k \in N}$. On pose $c_k(f) = \langle f, e_k \rangle$ et on définit :

$$\begin{aligned} \psi : H &\rightarrow l^2(N), f \mapsto (c_k(f))_{k \in N} \\ \psi &= (\langle f, e_1 \rangle, \langle f, e_2 \rangle, \dots, \langle f, e_k \rangle, \dots) \end{aligned}$$

L'injectivité de ψ en découle immédiatement. Par ailleurs, comme ψ est clairement linéaire, cela implique aussi sa continuité. Il reste à vérifier la surjectivité de ψ .

Soit donc $u = (u_k)_{k \in N}$ un élément quelconque de $l^2(N)$. Montrons l'existence d'un $f \in H$ tel que $\psi(f) = u$. Pour se faire, définissons la suite $(f_n)_{n \in N}$ d'éléments de H par :

$$\forall n \in N, f_n = \sum_{k=0}^n u_k e_k$$

Par orthonormalité des e_k , on a :

$$\forall n \in \mathbb{N}, \forall j \in \mathbb{N}^*; \|f_{n+j} - f_n\|^2 = \sum_{k=n+1}^{n+j} |u_k|^2 \leq \sum_{k=n+1}^{\infty} |u_k|^2 \xrightarrow{n \rightarrow \infty} 0$$

On en déduit que $(f_n)_{n \in \mathbb{N}}$ est de Cauchy dans H . Comme H est complet, $(f_n)_{n \in \mathbb{N}}$ a une limite f dans H , ce qui s'écrit toujours comme :

$$f = \sum_{k=0}^{\infty} u_k e_k$$

En vertu du théorème de l'unicité du développement, on a alors $u_k = c_k(f)$ et pour tout $k \in \mathbb{N}$, donc $\psi(f) = u$.

2.6 Espace de Sobolev

2.6.1 Dérivée faible

Dans ce qui suit, on considérera un intervalle ouvert (non vide) I de \mathfrak{R} , borné ou non, d'extrémités a et b ($-\infty < a < b < +\infty$); et on notera \bar{I} son adhérence dans \mathfrak{R} : $\bar{I} = [a, +\infty[$ si $I =]a, +\infty[$.

Avant de poursuivre, on se doit de définir certaines classes d'appartenances de fonctions que l'on va utiliser.

Donc, pour une fonction f définie sur un intervalle $I \subseteq \mathfrak{R}$, on dira que f est :

- de classe C^1 si f est dérivable sur $I \subseteq \mathfrak{R}$, et que sa dérivée f' est continu;

- de classe C^k si toutes les dérivées jusqu'à l'ordre k de f sur $I \in \mathfrak{R}$ existe, et que sa dérivée $f^{(k)}$ est continue;
- de classe C^∞ si $f \in C^k$ sur $I \in \mathfrak{R}$ pour tout $k \in \mathbb{N}$. Autrement dit, si f est indéfiniment différentiable.

De plus $D^k(I)$ désigne l'espace des fonctions de classe C^k à support compact dans $I \in \mathfrak{R}$,

Le point de départ de cette sous-section est la remarque suivante : si u est continûment différentiable sur $I =]a, b[$ et si $\varphi \in D^1(I)$ est une fonction de classe C^1 à support compact contenu dans I , alors, en intégrant par parties, on obtient :

$$\int_I u(t) \varphi'(t) dt = [u \varphi]_a^b - \int_I u'(t) \varphi(t) dt = - \int_I u'(t) \varphi(t) dt$$

Car, puisque le support de φ est un compact de $]a, b[$, et en vertu du fait que φ est nulle en dehors du compact $K \subset]a, b[$:

$$[u \varphi]_a^b = u(b) \varphi(b) - u(a) \varphi(a) = 0$$

cela conduit à la définition de la dérivée faible.

Définition 2.23 : On dit que $f \in L^p$, $1 \leq p < +\infty$, a une dérivée faible s'il existe $f' \in L^p$ tel que :

$$\int_I f \varphi' du = - \int_I f' \varphi du \quad \text{pour tout } \varphi \in D^1(I)$$

Proposition 2.8 : Si un tel $f' = v \in L^p$ existe, il est unique.

Démonstration : Pour cette démonstration, on utilise les notions de produit de convolution et d'unité rapprochée pour la convolution qui se définissent comme suit. Pour plus de détails et les démonstrations des résultats énoncés relatifs à la convolution, voir [Bracewell 1999] et [Samuelides 1990].

Définition 2.24 : Soient f et g sont deux fonctions définies sur les réels, leur produit de convolution $f * g$ est défini par :

$$(f * g)(x) = \int_{\mathfrak{R}} f(x-t)g(t)dt$$

Définition 2.25 : On appelle unité approchée pour la convolution toute suite (α_k) de fonctions à valeurs dans $[0, \infty[$ telle que :

1. Pour tout k , $\int_{\mathfrak{R}} \alpha_k(x)dx = 1$
2. $\int_{|x|>\delta} |\alpha_k(x)|dx \rightarrow 0$ pour tout $\delta > 0$ quand $k \rightarrow \infty$
3. Chaque α_k appartient à la classe C^∞

Dans cette preuve, on utilise aussi les résultats qui découlent des deux prochains théorèmes.

Théorème 2.2 : Si f est uniformément continue, $\alpha_k * f$ converge uniformément vers f .

Théorème 2.3 : Si $f \in L^p$, $\alpha_k * f$ converge vers f dans L^p

On peut déduire de ces deux théorèmes que si f , définie sur I , est localement intégrable, c'est-à-dire que f est intégrable sur tout intervalle fermé et borné contenu dans I , $\alpha_k * f$ est toujours une fonction de la classe C^∞ .

De plus, le produit de convolution fait de L^1 une algèbre où il n'y a pas d'élément unité, donc, il n'existe pas de fonction e telle que $e * f = f$ pour tout f . Dans ce cadre se sont les unités approchées qui jouent ce rôle au sens où :

$$\alpha_k * f \rightarrow f$$

On peut maintenant poursuivre la démonstration de l'unicité de $f' = v \in L^p$ en supposons qu'il existe $v_1, v_2 \in L^p$ tel que :

$$\int_I v_1 \varphi du = \int_I v_2 \varphi du \text{ pour tout } \varphi \in D^1(I)$$

Il suffit de montrer que $v = v_1 - v_2 \stackrel{p.p.}{=} 0$

Pour ce faire, soit $x_0 \in I$ et $r > 0$ tel que $[x_0 - 2r, x_0 + 2r] \in I$. Soit aussi $\rho_\varepsilon : \mathbb{R} \rightarrow \mathbb{R}, \varepsilon > 0$, des fonctions de classe C^∞ telles que $\text{supp}(\rho_\varepsilon) \subseteq [-r, r]$, forment une unité approchée pour la convolution. Lorsque $|x - x_0| \leq r$, la fonction suivante :

$$\varphi_\varepsilon : y \mapsto \rho_\varepsilon(x - y)$$

appartient à la classe C^∞ sur \mathfrak{R} et à support dans $[x_0 - 2r, x_0 + 2r] \in I$. Elle est donc dans $D^\infty(I) \subseteq D^1(I)$.

Posons maintenant :

$$\tilde{v} = v 1_{[x_0 - 2r, x_0 + 2r]}$$

Cela permet d'écrire :

$$\tilde{v} * \rho_\varepsilon(x) = \int_{x_0 - 2r}^{x_0 + 2r} v(y) \rho_\varepsilon(x - y) dy = \int_I v(y) \rho_\varepsilon(x - y) dy$$

$$\tilde{v} * \rho_\varepsilon(x) = \int_I (v_1(y) - v_2(y)) \varphi_\varepsilon(y) dy$$

D'autre part, comme $\tilde{v} \in L^1(\mathfrak{R})$, on a :

$$\tilde{v} * \rho_\varepsilon(x) \xrightarrow{\varepsilon \rightarrow 0} \tilde{v}$$

Donc $\tilde{v} = 0$ presque partout sur \mathfrak{R} . Cela signifie que $v = 0$ presque partout sur $[x_0 - 2r, x_0 + 2r]$. Comme c'est vrai pour tout $x_0 \in I$ et tout $r > 0$ tel que $[x_0 - 2r, x_0 + 2r] \subset I$, il en résulte que $v = 0$ presque partout sur I et donc l'unicité de $f' = v \in L^p$.

2.6.2 Espace de Sobolev

Définition 2.26 : Pour $1 \leq p < +\infty$, on définit l'espace de Sobolev $W^{1,p}$ par :

$$W^{1,p} = \{u \in L^p\}; u \text{ a une dérivée faible } u' \in L^p$$

On munit $W^{1,p}(I)$ de la norme

$$\|u\|_{W^{1,p}} = \|u\|_{L^p} + \|u'\|_{L^p}$$

On pourrait définir de même $W^{2,p}$, comme l'espace des $u \in L^p$ ayant une dérivée faible seconde, et plus généralement $W^{k,p}$ pour tout entier $k > 1$.

Pour $p = 2$, on note :

$$W^{1,2} = H^1$$

et on le munit du produit scalaire suivant :

$$\forall u, v \in H^1 \text{ on pose } \langle u, v \rangle = \langle u, v \rangle_{L^2} + \langle u', v' \rangle_{L^2}$$

La norme qui y est associée est :

$$\|u\|_{H^1} = \left(\|u\|_{L^2}^2 + \|u'\|_{L^2}^2 \right)^{1/2}$$

Proposition 2.9 : $W^{1,p}$ est un espace de Banach séparable. Il est réflexif pour $1 \leq p < +\infty$. De plus, H^1 est un espace de Hilbert séparable.

Démonstration : Si $(u_n)_{n \geq 1}$ est une suite de Cauchy dans $W^{1,p}$, alors $(u_n)_{n \geq 1}$ et $(u'_n)_{n \geq 1}$ sont des suites de Cauchy dans L^p . Donc

$$u_n \xrightarrow[n \rightarrow \infty]{} u$$

et

$$u'_n \xrightarrow[n \rightarrow \infty]{} v$$

Cela implique que, pour toute fonction $\varphi \in D^1$

$$\langle u, \varphi' \rangle = \lim_{n \rightarrow \infty} \langle u_n, \varphi' \rangle = - \lim_{n \rightarrow \infty} \langle u'_n, \varphi \rangle = - \langle v, \varphi \rangle$$

donc u a une dérivée faible et $u' = v$.

Posons maintenant :

$$T : \begin{cases} W^{1,p}(I) \rightarrow L^p(I) \times L^p(I) \\ u \mapsto T(u) = (u, u') \end{cases}$$

Comme T est une isométrie, que $L^p \times L^p$ est séparable et que pour $1 \leq p < +\infty$ il est réflexif, $W^{1,p}$ l'est aussi ([Brézis 2005]).

2.6.3 Théorème d'immersion

Théorème de Sobolev [Sobolev 1935] : Pour toute fonction $u \in W^{1,p}$, il existe une fonction \tilde{u} continue sur \bar{I} dont la restriction à I est égale à u . De plus :

$$\tilde{u}(y) - \tilde{u}(x) = \int_x^y u'(t) dt, \forall x, y \in I$$

On identifiera par la suite u à son représentant continu, de sorte que l'on peut écrire :

$$W^{1,p}(I) \subseteq C(\bar{I})$$

Démonstration : Fixons $x_0 \in I$ et posons :

$$u_0(x) = \int_{x_0}^x u'(t) dt, \forall x \in \bar{I}$$

Pour $p = 1$, il est clair que cela a un sens et que u_0 est continu, par le théorème de convergence dominée ([Colmez 2009]).

Pour $1 < p < \infty$ cette définition a un sens car si $J \subseteq \bar{I}$ est un intervalle de longueur finie, on a $L^p(J) \subseteq L^1(J)$. De plus, comme

$$u_0(x) - u_0(y) = \int_y^x u'(t) dt, \forall x, y \in I$$

la fonction $u_0 : I \rightarrow \mathfrak{R}$ est uniformément continue et par l'inégalité de Hölder on obtient :

$$|u_0(x) - u_0(y)| \leq |x - y|^{1/q} \|u'\|_p \leq |x - y|^{1/q} \|u\|_{W^{1,p}}$$

On peut donc la prolonger en une fonction uniformément continue sur \bar{I} .

Notons qu'en particulier u_0 est intégrable sur tout sous-intervalle $J \subseteq I$ de longueur finie.

Maintenant, pour toute fonction $\varphi \in D^1(I)$, on a :

$$\langle u_0, \varphi \rangle = \int_I \left(\int_{x_0}^x u'(t) dt \right) \varphi'(x) dx$$

$$\langle u_0, \varphi \rangle = - \int_a^{x_0} \left(\int_{x_0}^x u'(t) dt \right) \varphi'(x) dx + \int_{x_0}^b \left(\int_{x_0}^x u'(t) dt \right) \varphi'(x) dx$$

$$\langle u_0, \varphi \rangle \stackrel{\text{Fubini}}{=} - \int_a^{x_0} u'(t) \left(\int_a^t \varphi'(x) dx \right) dt + \int_{x_0}^b u'(t) \left(\int_t^b \varphi'(x) dx \right) dt$$

comme $\varphi \in D^1(I)$

$$\langle u_0, \varphi \rangle = - \int_a^{x_0} u'(t) (\varphi(t) - \varphi(a)) dt + \int_{x_0}^b u'(t) (\varphi(b) - \varphi(t)) dt$$

$$\langle u_0, \varphi \rangle = - \int u'(t) \varphi(t) dt = - \langle u', \varphi \rangle = \langle u, \varphi' \rangle$$

On peut ensuite remarquer que :

$$\{\varphi', \varphi \in D^1(I)\} = \left\{ \psi \in D^0(I); \int_I \psi(t) dt = 0 \right\}$$

et considérons aussi les formes linéaires :

$$J_1 : \begin{cases} D^0(I) \rightarrow \mathfrak{R} \\ \psi \mapsto \int_I \psi(t) dt \end{cases}$$

et

$$J_2 : \begin{cases} D^0(I) \rightarrow \mathfrak{R} \\ \psi \mapsto \int_I (u(t) - u_0(t)) \psi(t) dt \end{cases}$$

Les égalités précédentes entraînent que $\ker J_1 \subseteq \ker J_2$. Il existe donc

$c \in \mathfrak{R}$ tel que $J_2 = cJ_1$. Cela implique que :

$$\int_I (u(t) - u_0(t) - c) \varphi(t) dt = 0, \forall \varphi \in D^0(I)$$

C'est en particulier vrai pour toute fonction $\varphi \in D^1(I)$, et l'on a vu qu'alors

$u - u_0 - c = 0$ presque partout sur I .

Ainsi u est presque partout égal sur I à la fonction $u_0 + c$, qui est continue sur \bar{I} .

Maintenant que les bases théoriques nécessaires à la compréhension et à la mise en œuvre de l'analyse en composante principale et de l'analyse discriminante sont fixées, le chapitre suivant présentera les méthodes d'analyse de données dans le cadre où les données sont fonctionnelles.

Chapitre 3 : Présentations des méthodes d'analyses de données dans le cadre fonctionnel

3.1 Introduction

L'analyse de données fonctionnelles (ADF) est une extension de l'analyse des données classique, soit des données discrètes, appliquées à des observations qui sont, dans ce cadre, des fonctions. L'analyse de données fonctionnelles bien que récente à été tenté pour la première fois dans les travaux des mathématiciens Carl F. Gauss et Adrien-Marie Legendre sur la trajectoire des comètes [Gauss 1809] et [Legendre 1805]. Il a par contre fallu attendre plusieurs années pour que les méthodes se développent réellement. À partir de 1945, plusieurs textes influents, de nature plus technique, ont été publiés ce qui a permis à quelques mathématiciens et statisticiens en France de parler pour la première fois de l'analyse en composantes principales de données fonctionnelles. Citons parmi ceux-ci, les travaux de Jean-Claude Deville qui a considéré une ACP sur des données fonctionnelles prenant des valeurs dans L^2 [Deville 1974] et aussi les travaux de Jacques Dauxois et Alain Pousse qui ont formalisé la version fonctionnelle de l'ACP [Dauxois *et al.* 1976]. Les objectifs de l'analyse de données fonctionnelles sont essentiellement les mêmes que les méthodes statistiques traditionnelles. Nous pouvons les résumer par :

- Représenter les données pour faciliter les étapes futures d'analyse
- Faire ressortir les caractéristiques prédominantes des données
- Étudier les tendances et les variations qui se dégagent des données
- Expliquer les variations des variables dépendantes à l'aide des variables indépendantes

- Comparer au moins deux ensembles de données pour comprendre leurs différences et leurs similitudes

3.2 Lien entre la théorie et la pratique

Avant de poursuivre, il est essentiel de faire le lien entre le chapitre deux qui présente le cadre théorique et ce présent chapitre qui présente les aspects pratique de la réalisation de l'analyse en composantes principales et l'analyse discriminante de données fonctionnelles. Le texte qui suit est fortement inspiré de [Besse 1991].

On commence par définir le contexte dans lequel nous allons aborder la question.

Soit H un espace hilbertien réel, H' l'espace dual associé à H (dual topologique) ou, en d'autres termes, l'espace des fonctionnelles linéaires continues et β_H la σ -algèbre des Boreliens de H .

Soit également Y une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathcal{F}, \mu)$ et à valeurs dans (H, β_H) . On note par $L^2(\mu)$ l'espace $L^2(\Omega, \mathcal{F}, \mu)$ des variables aléatoires réelles de carré intégrable, muni du produit scalaire usuel et l'on désigne les isométries canoniques entre H et H' et entre $L^2(\mu)$ et $L^2(\mu)'$ respectivement par M et D .

Enfin, on supposera dans la suite que Y est de second ordre ou en d'autres termes :

$$E \|Y\|_H^2 < \infty \text{ ce qui signifie que } Y \in L^2(\Omega, \mathcal{A}, \mu; H, \beta_H) \text{ noté } L_H^2(\mu).$$

En vertu du théorème de représentation de Riesz des formes linéaires (démontré dans [Brezis 2005]) on sait qu'il existe un élément de H , noté $E(Y)$ (l'espérance), et un opérateur V de l'espace des formes linéaires qui sont de carrées intégrables de H' dans H , noté $\mathbf{L}_H^2(H', H)$, (la covariance) définis de manière unique par les relations suivantes :

$$\begin{cases} \forall u \in H', E \langle Yu \rangle_{HH'} = \langle E(Y), u \rangle_{HH'} \\ \forall u_1, u_2 \in H' \\ E [\langle Y - E(Y), u_1 \rangle_{HH'} \cdot \langle Y - E(Y), u_2 \rangle_{HH'}] = \langle Vu_1, u_2 \rangle_{HH'} \end{cases}$$

où $\langle Y, u \rangle_{HH'} = Y(u)$ est le produit de dualité entre H et H' (voir [Schwartz 1979]).

Avant de poursuivre, on se doit d'abord d'introduire la notion d'opérateur et quelques caractéristiques qui peuvent lui être associées.

Définition 3.1 : En mathématique, un opérateur linéaire, O , est une application entre deux espaces vectoriels, E, F , et il est noté de la façon suivante : $O : E \rightarrow F$.

On rappelle aussi qu'en dimension infinie, un espace de Hilbert, H , est séparable si et seulement si il admet une base orthonormée complète, $\{e_i\}, i \in \mathbb{Z}$.

Définitions 3.2 : Un opérateur, $O : E \rightarrow F$, est dit :

1. positif, si le produit scalaire $\langle Ox, Oy \rangle$ est positif pour tout $x, y \in E$.
2. borné, s'il existe un réel, M , strictement positif, $M > 0$, tel que : $\|Ox\| \leq M$ pour tout $x \in E$.
3. à trace, si $\|O\| = \sum_{i \in \mathbb{Z}} \|Oe_i\| < \infty$. Dans ce cas, on pose : $Tr(O) = \sum_{i \in \mathbb{Z}} \langle Oe_i, e_i \rangle$.
4. compact, si l'adhérence dans F de l'image par O de la boule unité de E , $O(B_E)$, est compacte dans l'espace F .
5. nucléaire, s'il est un opérateur compact et à trace.
6. auto-adjoint, si :

$$\langle Ou, v \rangle = \langle u, Ov \rangle \text{ pour tout } u, v \in E$$

7. de Hilbert-Schmidt, sous les conditions suivantes : Soit H un espace de Hilbert séparable de dimension infinie et soit $T \in L(H)$ un opérateur de H . On dit que T est un opérateur de Hilbert-Schmidt, s'il existe une base hilbertienne $(e_n)_{n \in \mathbb{N}}$ telle que :

$$\sum_{n=1}^{\infty} \|Te_n\|^2 < \infty$$

On pose alors $X = Y - E(Y)$ qui est une variable aléatoire hilbertienne réelle centrée. L'opérateur de covariance est borné, $\|Vu\|_H \leq \|u\|_H$, positif, $\langle Vu, u \rangle_{HH} > 0$,

et admet pour trace $E \|X\|_H^2 < \infty$ sous l'hypothèse que $E \|Y\|_H^2 < \infty$ [Brézis 2005]. On peut donc dire que $E \|X\|_H^2$ est un opérateur nucléaire.

De plus, si H_2 désigne un autre espace hilbertien, pour toute application linéaire continue $\Pi : H \rightarrow H_2$, $\Pi(Y)$ est une variable aléatoire hilbertienne élément de $L_{H_2}^2(\mu)$. Alors $\Pi(Y)$ admet pour espérance $\Pi(E(Y))$ et pour covariance l'opérateur nucléaire :

$$V_2 = \Pi V \Pi^T$$

Il existe de nombreuses façons équivalentes de définir une analyse en composantes principales ou une analyse discriminante et elles se ramènent toutes, lorsque l'hypothèse $E \|Y\|_H^2 < \infty$ est vérifiée, à l'analyse spectrale d'un opérateur compact, La procédure adoptée ici est tirée de [Besse 1991] et est une démarche pas à pas classique. Le développement est présenté spécialement pour l'ACP car, comme cela a été écrit précédemment, l'analyse discriminante peut être vue comme une analyse en composantes principales des centres de gravité des classes sous certaines conditions qui n'affectent en rien les hypothèses de base.

Comme tout élément $u \in H'$ est continue, alors $u(X(\bullet))$ qui n'est autre que $\langle X(\bullet), u \rangle_{HH'}$, est une variable aléatoire réelle définie sur (Ω, F, μ) . De plus, elle est intégrable si la condition suivante est satisfaite :

$$E \|X\|_H^2 < \infty$$

car dans ce cas on a :

$$\left| \langle X(\bullet), u \rangle_{HH} \right| \leq \|u\|_H \cdot \|X(\bullet)\|_H$$

Notons par U l'application linéaire de H' dans $L^2(\mu)$ définie par :

$$U(u) = \langle X(\bullet), u \rangle_{HH} ,$$

Proposition 3.1 : L'opérateur U défini ci-dessus est un opérateur de Hilbert-Schmidt ainsi que l'application U^T définie par :

$$\forall f \in L^2(u)', U^T(f) = E(XD^{-1}f)$$

où $D : L^2(u) \rightarrow L^2(u)'$.

De plus, la covariance V de X peut s'écrire comme $U^T D U$.

Démonstration : Si $\{\beta_i\}_{i \in I}$ désigne une base orthonormée de H' , cela implique que :

$$\sum_{i \in I} \|U\beta_i\|_{L^2(u)}^2 = \sum_{i \in I} E \langle X, \beta_i \rangle_{HH}^2 = E \|X\|_H^2 \leq \infty$$

Alors, d'après un théorème de Guelfand et Vilenkin [Guelfand et al. 1967], l'opérateur U est un opérateur de Hilbert-Schmidt ainsi que son transposé qui vérifie que pour tout $f \in L^2(u)'$ et pour tout $u \in H'$:

$$\begin{aligned}
\langle U(u), f \rangle_{L^2(u)L^2(u)} &= \langle U(u), D^{-1}f \rangle_{L^2(u)} = E(\langle X, u \rangle_{HH}, D^{-1}f) \\
&= E\langle XD^{-1}f, u \rangle_{HH} = \langle E(XD^{-1}f), u \rangle_{HH} \\
&= \langle U^T f, u \rangle_{HH}.
\end{aligned}$$

De plus, $\forall u_1, u_2 \in H'$:

$$\begin{aligned}
\langle U^T DU u_1, u_2 \rangle_{HH} &= \langle E(\langle X, u_1 \rangle_{HH}, X), u_2 \rangle_{HH} \\
&= E(\langle X, u_1 \rangle_{HH}, \langle X, u_2 \rangle_{HH}) \\
&= \langle Vu_1, u_2 \rangle_{HH}.
\end{aligned}$$

Ce qui montre également que V est nucléaire comme produit de deux opérateurs de Hilbert-Schmidt.

Les opérateurs qui ont été présentés sont illustrés à la figure 11. Elle présente le diagramme commutatif ou le schéma de dualité existant entre les espaces définis ci-dessus.

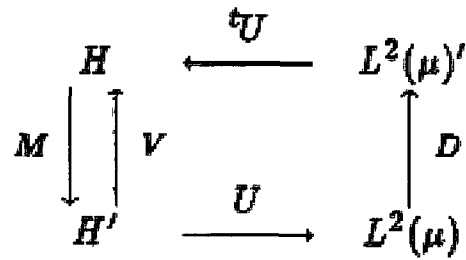


Figure 11: Schémas de dualité

On peut maintenant définir formellement l'analyse en composantes principales.

Définition 3.3 : On appelle analyse en composantes principales de Y dans H la recherche dans H' d'un élément u de norme 1 tel que $U(u)$ soit de norme maximale dans $L^2(\mu)$ et sous contrainte d'orthonormalité dans H' .

L'orthonormalité est obtenue à l'aide du processus défini de façon récurrente comme suit :

Si les k premiers éléments u_1, \dots, u_k ont pu être obtenus, alors u_{k+1} est solution de :

$$\max \left\{ \|U(u)\|_{L^2(\mu)} \mid u \in H', \|u\|_{H'} = 1, \forall i \leq k : (u_i \mid u)_{H'} = 0 \right\}$$

Proposition 3.2 : L'analyse en composantes principales de Y dans H conduit à l'analyse spectrale de l'opérateur nucléaire, auto-adjoint et positif MV .

Démonstration : En remplaçant dans l'équation précédente la norme par son carré, le problème s'écrit pour la première étape :

$$\begin{aligned}
& \max \left\{ \left\| U(u) \right\|_{L^2(u)}^2 \mid u \in H', \|u\|_{H'} = 1 \right\} \\
&= \max \left\{ \langle Vu, u \rangle_{HH'} \mid u \in H', \|u\|_{H'} = 1 \right\} \\
&= \max \left\{ \langle MVu, u \rangle_{H'} \mid u \in H', \|u\|_{H'} = 1 \right\}
\end{aligned}$$

Comme V est positif, nucléaire et M positif, borné, le produit MV est aussi positif et nucléaire. La solution est le vecteur propre u_1 de MV associé à la plus grande valeur propre λ_1 . De plus, d'après les propriétés spectrales d'un opérateur compact positif, le maximum est atteint pour les autres étapes avec u_k vecteur propre de MV associé au $k^{ième}$ élément de la suite décroissante $\{\lambda_i \mid i \in I\}$ des valeurs propres [Besse 1991].

Donc, la démonstration de cette proposition prouve que la résolution d'un problème d'analyse en composantes principales et d'analyse discriminante se résume à la recherche des valeurs et des vecteurs propres dans le cas où la variable aléatoire respecte l'hypothèse $E \|Y\|_H^2 < \infty$. Donc, pour la partie suivante de ce chapitre, on supposera qu'elle est vérifiée. De plus, on retrouve au chapitre suivant, la validation de cette hypothèse pour l'exemple avec les données de prises alimentaires.

3.3 Analyse en composantes principales pour des données fonctionnelles.

Commençons avec le même jeu de données classiques du chapitre un.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \text{où } x_{ij} \in \mathfrak{R} \quad \forall i : 1, 2, \dots, n; j : 1, 2, \dots, p$$

Plusieurs bases, telles les séries de Fourier, les polynômes orthogonaux, les splines, les B-splines, les ondelettes, etc., sont disponibles pour passer de données classiques x_{ij} aux données fonctionnelles $x_i(s)$ avec $i : 1, 2, \dots, n$ et $s \in [a, b]$ un intervalle de \mathfrak{R} . On choisira cette dernière en fonction des caractéristiques de données brutes et des besoins des analyses à effectuer. On présentera, au chapitre suivant, ce qui a motivé le choix de la méthode pour l'étude de cas sur les données de prises alimentaires porcines. Le vecteur $X(s)$ contenant les données fonctionnelles peut s'écrire comme :

$$X(s) = (x_1(s) \quad x_2(s) \quad \dots \quad x_n(s))^T$$

Donc, le problème dans le cadre de données fonctionnelles, consiste à trouver un ensemble de fonctions $\xi_\alpha(s)$ où $\alpha : 1, 2, \dots, q$, mutuellement orthogonales, de norme 1 ($\xi_\alpha^T \xi_\alpha = 1$), ce qui est l'équivalent des vecteurs propres dans le cadre classique. Chacune des fonctions $\xi_\alpha(s)$ doit maximiser la quantité suivante :

$$\frac{1}{n} \sum_{i=1}^n \left(\int_a^b \xi_\alpha(s) x_i(s) ds \right)^2 = \frac{1}{n} \sum (\langle x_i, \xi_\alpha \rangle)^2 = \frac{1}{n} (c_\alpha(x_i))^2$$

où $c_\alpha(x_i)$ n'est autre que la coordonnée de la projection orthogonale de x_i sur ξ_α et, en ce sens, l'expression ci-dessus correspond bien à la maximisation de l'inertie.

En introduisant la fonction de covariance $v(s, t)$, $s, t \in [a, b]$, définie de la façon suivante :

$$v(s, t) = \frac{1}{n} \sum_{i=1}^n x_i(s) x_i(t)$$

on peut réécrire l'équation à maximiser comme suit :

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left(\int_a^b \xi_\alpha(s) x_i(s) ds \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left(\int_a^b \xi_\alpha(s) x_i(s) ds \int_a^b \xi_\alpha(t) x_i(t) dt \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\int_a^b \xi_\alpha(s) \left(\int_a^b \xi_\alpha(t) x_i(s) x_i(t) dt \right) ds \right) \\ &= \int_a^b \xi_\alpha(s) \left[\int_a^b v(s, t) \xi_\alpha(t) dt \right] ds \end{aligned}$$

De plus, on sait, par [Aubin 2000], que cette maximisation est équivalente à trouver les éléments propres de $v(s, t) = \frac{1}{n} X^T(s) X(t)$ que l'on nomme opérateur de covariance. Donc, cela nous ramène au principe de base de l'ACP.

Sans perte de généralité, on peut supposer que nous avons des données fonctionnelles centrées pour résoudre le problème des éléments propres. On peut également écrire les données fonctionnelles, $x_i(s)$, comme des

combinaisons linéaires de fonctions de base de l'espace fonctionnel. Cela permet d'affirmer que chacune des composantes principales fonctionnelles peut être représentée par une fonction. Donc, chacune des observations, est aussi une fonction continue, combinaison linéaire des fonctions de base $\phi_k(s)$ avec $k = 1, 2, \dots, q, \dots$ et où chacune de ces fonctions est pondérée par un coefficient c_{ik} . Ainsi, on a :

$$x_i(s) = \sum_{k=1}^{\infty} c_{ik} \phi_k(s)$$

On peut maintenant écrire le vecteur $X(s)$ des données fonctionnelles comme étant :

$$X(s) = C\Theta(s)$$

avec C la matrice n par q des coefficients de combinaisons linéaires c_{ik} et $\Theta(s) = \phi_k(s)$ le vecteur des q fonctions de base.

La fonction $v(s, t)$ de covariance peut être réécrite comme :

$$v(s, t) = \frac{1}{n} X^T(s) X(t) = \frac{1}{n} \Theta^T(s) C^T C \Theta(t)$$

On peut écrire chacune des fonctions propres $\xi_\alpha(s)$ à l'aide des fonctions de base $\phi_k(s)$. Ce qui nous donne :

$$\xi_\alpha(s) = \Theta^T(s) b_\alpha$$

où $b_\alpha = [b_{\alpha 1}, b_{\alpha 2}, \dots, b_{\alpha q}]^T$ sont les coefficients de la combinaison linéaire. Alors en utilisant V tel que défini précédemment, on obtient :

$$V\xi_\alpha(s) = \int_a^b v(s,t)\xi_\alpha(t)dt$$

$$V\xi_\alpha(s) = \frac{1}{n} \int_a^b \Theta^T(s)C^T C\Theta(t)\xi_\alpha(t)dt = \frac{1}{n} \int_a^b \Theta^T(s)C^T C\Theta(t)\Theta^T(t)b_\alpha dt$$

$$V\xi_\alpha(s) = \frac{1}{n} \Theta^T(s)C^T CWb_\alpha$$

où W est la matrice $q \times q$ des coefficients $w_{kl} = \int_a^b \phi_k(s)\phi_l(s)ds$ avec $k, l = [1, 2, \dots, q]$.

On se ramène au problème de l'analyse en composantes principales fonctionnelles, c'est-à-dire, un problème de valeur propre, car :

$$V\xi_\alpha(s) = \frac{1}{n} \Theta^T(s)C^T CWb_\alpha$$

$$V\xi_\alpha(s) = \lambda_\alpha \xi_\alpha(s) = \lambda_\alpha \Theta^T(s)b_\alpha$$

où λ_α est la valeur propre associée à la fonction propre $\xi_\alpha(s)$ et cela revient à dire que :

$$\forall s \in [a, b] \exists \text{ au moins une solution tel que } \frac{1}{n} C^T CWb_\alpha = \lambda_\alpha b_\alpha$$

Par ailleurs, la contrainte :

$$\int_a^b \xi_\alpha^2(s)ds = 1 \quad \forall \alpha$$

permet d'écrire que :

$$\int_a^b b_\alpha^T \Theta(s) \Theta^T(s) b_\alpha ds = b_\alpha^T W b_\alpha = 1$$

Pour être conforme à la notation précédente, posons μ_α un vecteur de norme 1 tel que $\mu_\alpha = W^{1/2} b_\alpha$ et donc :

$$\frac{1}{n} C^T C W b_\alpha = \lambda_\alpha b_\alpha \equiv \frac{1}{n} (W^{1/2} C^T C W^{1/2}) \mu_\alpha = \lambda_\alpha \mu_\alpha$$

En résumé, l'analyse en composantes principales fonctionnelles consiste à calculer les éléments propres, soient les $\mu_\alpha = \mu_1, \mu_2, \dots, \mu_{n-1}$ et les $\lambda_\alpha = \lambda_1, \lambda_2, \dots, \lambda_{n-1}$ de la matrice $\frac{1}{n} (W^{1/2} C^T C W^{1/2})$. Par la suite, il faut calculer chacune des composantes principales fonctionnelles $\xi_\alpha(s)$ avec $\alpha = 1, 2, \dots, q$. Notons qu'il y a que $n-1$ valeurs propres ou fonctions propres, car, contrairement au cas multidimensionnel classique où l'on travaille dans \mathfrak{R}^p ou \mathfrak{R}^n , les données fonctionnelles sont dans \mathfrak{R}^∞ . Par contre, comme les $x_i(s)$ ne sont pas linéairement indépendants, la matrice V est de rang $n-1$ et donc, il y aura $n-1$ valeurs propres différentes de zéro.

3.4 Analyse discriminante pour des données fonctionnelles.

Nous nous retrouvons toujours avec les mêmes données de base, soient :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \text{où } x_{ij} \in \mathbb{R} \quad \forall i : 1, 2, \dots, n; j : 1, 2, \dots, p$$

que l'on a préalablement transformées en données fonctionnelles :

$$X(s) = (x_1(s) \quad x_2(s) \quad \dots \quad x_n(s))^T$$

Supposons maintenant que les données soient séparées en deux catégories mutuellement exclusives par une variable indicatrice $\partial_i = 1$, où $i = 1, 2, \dots, n$, si la $i^{\text{ème}}$ observation fait partie de la première catégorie et $\partial_i = -1$ sinon. Dans ce cadre général, le but de l'analyse discriminante fonctionnelle (ADF) est de trouver la fonction $v(s)$ qui prédit le mieux la variable ∂_i pour chacun des individus de notre échantillon ou pour un nouvel ensemble d'individus.

Cette fonction peut s'écrire de la façon suivante :

$$v(s) = (v_1(s), v_2(s), \dots, v_p(s))$$

où chacun des $v_k(s)$ représente la valeur que prend la fonction discriminante sur chacune des dimensions de l'espace des données.

La prédiction, $\hat{\partial}_i$, pour ∂_i peut donc s'écrire de la façon suivante :

$$\hat{\partial}_i = \int_a^b X(s)v(s)ds$$

$$\hat{\partial}_i = \int_a^b (x_i(s)v_1(s) + x_i(s)v_2(s) + \dots + x_i(s)v_p(s))ds$$

Par la suite, il faut regarder si la valeur discriminante prédite, $\hat{\partial}_i$, est plus grande ou plus petite que la valeur critique C qui sert à départager les deux groupes. Mais, la vraie question est comment nous pouvons déterminer $v(s)$?

Supposons pour un moment que l'on est dans le cas où X est une matrice contenant les vecteurs des données à la place des fonctions. Le problème décrit précédemment revient à trouver un vecteur v et une constante C qui nous permettent de déterminer, le mieux possible, la catégorie de provenance d'un vecteur X_i en utilisant une inéquation de la forme $v'X_i > C$. C'est ce que l'on appelle communément l'analyse discriminante linéaire.

Comme présentée au chapitre 1, elle se résume à trouver le vecteur v qui minimise le rapport entre la variance interclasse et la variance intra-classe.

On pose \bar{X}^1 et \bar{X}^2 les moyennes des deux échantillons identifiés par la variable ∂ et \hat{S} la matrice de variance estimée. À partir de cela, l'analyse discriminante linéaire peut s'écrire de la façon suivante :

$$v = \hat{S}^{-1}(\bar{X}^2 - \bar{X}^1)$$

et

$$C = \frac{1}{2}v^T(\bar{X}^2 + \bar{X}^1)$$

Dans le cas fonctionnel, il n'est pas de mise d'appliquer telle quelle une extension de cette méthode en raison du problème de sur-dimensionnalité. Pour pallier ce problème, on peut régulariser les données [Gareth et Hastie 2001]. Une façon simple de le faire est de représenter les données et les $v(s)$ dans une certaine base où l'on ne considère qu'un nombre fini de termes. Dans le cas présent, l'ACP donne un espace qui, par l'ordre des composantes, préserve le plus possible la variabilité de l'échantillon. Pour cette raison, on peut utiliser comme nouvelle base les J premiers axes principaux déjà calculés.

Il est maintenant possible d'appliquer la méthode régulière d'analyse discriminante linéaire à cette nouvelle matrice Z de dimension $n \times J$ ($z_{i,k}, i = 1, 2, \dots, n; k = 1, 2, \dots, j$).

Cela donne un nouveau vecteur v de dimension J . La variable prédite, $\hat{\delta}_i$, peut alors s'écrire de la façon suivante :

$$\hat{\delta}_i = \sum_{k=1}^J v_k z_{i,k}$$

En utilisant les propriétés des composantes principales, on peut écrire $\hat{\delta}_i$ en termes de fonctions en écrivant les $z_{i,k}$ comme suit :

$$z_{i,k} = \int_a^b (x_i(s) \xi_j^1(s) + x_i(s) \xi_j^2(s) + \dots + x_i(s) \xi_j^p(s)) ds$$

ce qui donne l'équation suivante pour les $\hat{\delta}_i$:

$$\hat{\delta}_i = \sum_{j=1}^J v_j \int_a^b (x_i(s) \xi_j^1(s) + x_i(s) \xi_j^2(s) + \dots + x_i(s) \xi_j^p(s)) ds$$

Si on écrit les $v(s)$ de la manière suivante :

$$\begin{bmatrix} v_1(s) \\ v_2(s) \\ \vdots \\ v_p(s) \end{bmatrix} = \sum_{j=1}^J v_j \begin{bmatrix} \xi_j^1(s) \\ \xi_j^2(s) \\ \vdots \\ \xi_j^p(s) \end{bmatrix}$$

Ceci ramène au problème classique défini plus haut :

$$\hat{\partial}_i = \int_0^1 (x_i(s)v_1(s) + x_i(s)v_2(s) + \dots + x_i(s)v_p(s))ds$$

3.5 Résumé

Grâce à ce chapitre, nous connaissons maintenant le lien entre la théorie de l'analyse fonctionnelle et les méthodes statistiques multidimensionnelles présentées dans le cadre classique. Il a aussi été expliqué pourquoi l'analyse en composantes principales et l'analyse discriminante se résument en la recherche des vecteurs et des valeurs propres sous certaines hypothèses. Donc, il est maintenant possible de passer à l'application de ces dites méthodes aux données de prises alimentaires quotidiennes porcines.

Chapitre 4 : L'analyse des données

4.1 Introduction

On a mentionné précédemment que l'on appliquera les méthodes présentées dans le chapitre 3 à des données de prises alimentaires porcines dans le but de détecter s'il y a un lien entre la consommation quotidienne de nourriture et le fait, pour un porc, d'être malade ou non. S'ajoutant à la description statistique habituelle, une brève section concernant le fait que les données proviennent d'animaux vivants est ajoutée et intitulée : méthodologie expérimentale. Pour de plus amples détails sur le sujet, voir [Pomar *et al.* 2004]. Cette section a pour but de présenter les différents traitements et mesures que les porcs ont subis pendant la durée de l'expérience. Cette méthode permettra aux différents chercheurs dans le domaine de la production porcine de comprendre comment les porcs ont été traités et de savoir si cette étude pourra avoir des effets sur la production à grande échelle ou hors d'un centre de recherche. De plus, la description de cette méthodologie démontrera que la majorité des facteurs extérieurs pouvant avoir un effet sur les données de prise alimentaire ont été contrôlés.

4.2 Méthodologie expérimentale

4.2.1 Animaux et régie

Deux cent dix porcelets mâles castrés d'un poids approximatif de 25 kg (± 3 kg) ont été transportés au Centre de recherche sur le bovin laitier et le porc de Lennoxville. Tous les animaux étaient en excellente santé avant leur départ pour Lennoxville. Les porcelets ont été gardés en groupe dès leur arrivée; ils ont été alimentés avec une moulée de début de croissance commerciale médicamentée de haute performance. Durant la phase expérimentale, l'eau et la nourriture ont

été disponibles sans restriction. Une semaine après leur arrivée, les porcelets ont été à nouveau pesés et identifiés. Les 208 porcelets des classes centrales de poids ont été assignés aux quatre traitements selon un dispositif aléatoire. Seulement 12 porcs par parc ont subis cette expérience. Aussi, les huit parcs ont été attribués aux traitements et ce, aléatoirement. Les 18 porcelets restants ont été vendus quatre semaines après le début de l'expérience, car ils n'ont pas été utilisés comme remplacement durant l'expérience. Les porcs ont été maintenus à une température approximative de 22 degrés Celsius au début de l'expérience et elle a été diminuée de 0,5 degré Celsius par semaine. Les animaux ont été gardés en accord avec les recommandations du code des bonnes pratiques [Agriculture and Agri-Food Canada, 1993] et le guide du Canadian Council on Animal Care (1993).

Les animaux ont été abattus dans un abattoir commercial une semaine après la fin de l'expérience. La température de chaque section de la porcherie et l'humidité relative ont été mesurées deux fois par jour, la première au début de la journée et la deuxième à 16:00 heures. Tous les événements irréguliers tels la présence de selles liquides, fièvre, traitements, etc. ont été notés. Dans notre cas, la présence de fièvre sera l'indicateur de l'évolution de l'état de santé d'un porc. La présence de membres du personnel extérieurs et de visiteurs sera aussi notée. L'entrée de visiteurs sera limitée.

4.2.2 Aliments

Les phases d'alimentation ont duré 28 jours chacune ce qui correspondra à des poids approximatifs de 30 à 55 kg, de 55 à 85 kg et de 85 à 118 kg. Tous les aliments servis aux porcs ont été obtenus par l'amalgame des mêmes deux pré-mélanges. Ces combinaisons ont été formulées sans imposer qu'ils soient complets et à densité optimale; le tout par la méthode de formulation récemment développée [Letourneau-Montminy et al. 2005]. Les aliments attribués aux porcs

nourris selon un système d'alimentation en phases quotidiennes seront obtenus par mélange de ces deux pré-mélanges; à chaque jour, les porcs ont reçus un aliment complet dont la concentration en nutriments correspondait à trois niveaux de nutriments différents. La concentration de nutriments de ces deux pré-mélanges a été telle qu'ils puissent transmettre tous les aliments nécessaires à la réalisation de ce projet. Pour le traitement reproduisant les conditions habituelles d'alimentation canadiennes, soit la régie d'alimentation à trois phases (traitement 3P), les trois aliments nécessaires ont été obtenus par le mélange des pré-mélanges. Cela pour que la concentration en nutriments soit en accord avec les pratiques courantes canadiennes en suivant les recommandations de spécialistes chevronnés de la nutrition du porc. Pour les traitements reproduisant un système d'alimentation à phases quotidiennes, le premier traitement (traitement MC1) a fourni, pour l'ensemble de la période de croissance, un apport de lysine similaire à celui du traitement à trois phases (3P) [Pomar 2007]. Le deuxième niveau a fourni un apport en lysine qui correspondra aux besoins nutritionnels calculés pour chaque journée (MC2). Finalement, un dernier traitement fournira un apport en lysine (MC3) qui a été d'autant déficitaire que le traitement MC1 était excédentaire; autrement dit, $\text{lysMC3} = \text{lysMC2} - (\text{lysMC1} - \text{lysMC2})$.

Tous les aliments et pré mélanges ont été formulés avec une composition fixe de 15% de blé dur (11 % protéine brute, noté PB). Le tourteau de soja (48% PB) et la farine de poisson ont été utilisés comme principales sources de protéine alors que le maïs et le blé comme principales sources d'énergie. Les acides aminés de synthèse ont été utilisés pour maximiser l'équilibre des acides aminés de la protéine de la ration. La composition en minéraux et vitamines a été formulée de façon à satisfaire les besoins des animaux à haut potentiel de dépôt protéique.

4.3 Données

La base de données contient 177 porcs qui ont été séparés en 16 parcs. Pour chacun d'eux, on dispose de 83 observations de consommation quotidienne de nourriture (en kg/jour). Par contre, la prise des mesures n'a pas débutée simultanément pour chacun des parcs. Le tableau 13 présente les différentes dates de début et de fin pour chacun des parcs.

Tableau 13 : Dates de début et de fin de la prise de mesure pour les 16 différents parcs

Parcs	Nombre de porc	Date du début du traitement	Date de fin du traitement
1	11	8 novembre 2006	29 janvier 2007
2	11	9 novembre 2006	30 janvier 2007
3	12	10 novembre 2006	31 janvier 2007
4	12	11 novembre 2006	1 février 2007
5	11	15 novembre 2006	5 février 2007
6	12	16 novembre 2006	6 février 2007
7	12	17 novembre 2006	11 février 2007
8	11	21 novembre 2006	12 février 2007
9	12	29 novembre 2006	19 février 2007
10	11	30 novembre 2006	20 février 2007
11	11	1 décembre 2006	21 février 2007
12	12	2 décembre 2006	22 février 2007
13	9	6 décembre 2006	26 février 2007
14	10	7 décembre 2006	27 février 2007
15	10	8 décembre 2006	28 février 2007
16	10	9 décembre 2006	1 mars 2007

Donc, on doit poser les hypothèses suivantes :

1. La date de début de traitement n'a aucun effet sur la prise alimentaire des porcs;
2. Tous les parcs sont identiques et indépendants;
3. Les 15 porcs qui ont été retirés de la base de données l'ont été pour des raisons autres que celles que l'on désire étudier.

Le tableau 14 présente les moyennes par semaine de la consommation quotidienne, en kilogramme, pour les différents parcs. Les nombres en caractères gras sont les minimums et les maximums pour chacune des 12 semaines.

Le tableau 15 présente les écarts types par semaine de la consommation quotidienne, en kilogramme, pour les différents parcs.

La combinaison de ces deux tableaux permet de conclure qu'il n'y pas de différence significative entre les différents parcs. Aussi, pour la suite de l'analyse on ignorera de cette variable comme le suggère l'hypothèse 2.

Tableau 14: Moyenne hebdomadaire de la consommation quotidienne (kg) pour les différents parcs

Parc	Semaine											
	1	2	3	4	5	6	7	8	9	10	11	12 ²
1	1,8	2,2	2,4	2,5	2,7	2,7	2,7	2,8	2,9	3,0	3,1	3,0
2	1,7	2,0	2,2	2,4	2,5	2,6	2,6	2,7	3,0	3,1	3,3	3,3
3	1,8	2,0	2,2	2,2	2,4	2,5	2,6	2,7	2,9	3,1	3,1	3,0
4	1,9	2,0	2,3	2,4	2,4	2,5	2,6	2,7	2,8	2,9	3,1	3,1
5	1,8	2,1	2,4	2,6	2,6	2,8	2,8	2,8	3,2	3,3	3,6	3,4
6	2,2	2,4	2,5	2,8	2,8	2,8	2,9	2,9	2,9	2,8	2,9	2,8
7	1,9	2,1	2,2	2,4	2,5	2,5	2,5	2,7	2,8	2,8	2,9	2,9
8	1,8	2,0	2,2	2,4	2,5	2,6	2,7	2,8	3,0	3,2	3,3	3,3
9	1,8	1,7	1,9	2,0	2,1	2,2	2,3	2,5	2,9	3,1	3,2	3,2
10	1,8	1,8	2,1	2,1	2,2	2,3	2,3	2,4	2,7	2,8	2,8	2,8
11	1,7	1,9	2,0	2,2	2,3	2,5	2,6	2,6	3,0	2,9	3,1	3,0
12	1,8	2,0	1,9	2,1	2,3	2,4	2,5	2,6	2,8	2,8	2,9	2,9
13	1,9	2,1	2,2	2,3	2,5	2,4	2,5	2,6	2,8	3,0	3,1	3,0
14	1,8	1,9	1,9	2,2	2,3	2,5	2,6	2,8	2,8	2,9	3,1	3,1
15	1,6	2,0	2,0	2,1	2,2	2,4	2,5	2,7	2,8	2,8	2,9	2,8
16	1,7	1,9	2,0	2,1	2,2	2,3	2,2	2,4	2,5	2,7	2,9	2,7

² La semaine 12 ne comporte que 6 jours

Tableau 15: Écart-type hebdomadaire de la consommation quotidienne (kg) pour les différents parcs

Parc	Semaine											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0,4	0,4	0,7	0,4	0,5	0,5	0,4	0,6	0,5	0,6	0,8	0,9
2	0,4	0,4	0,5	0,5	0,7	0,5	0,5	0,6	0,7	1,0	0,7	1,1
3	0,4	0,5	0,3	0,4	0,4	0,4	0,4	0,5	0,4	0,6	0,6	0,8
4	0,3	0,3	0,3	0,5	0,5	0,5	0,6	0,7	0,6	0,7	1,0	1,2
5	0,6	0,6	0,6	0,6	0,5	0,5	0,8	0,8	1,7	1,7	2,0	1,7
6	0,3	0,5	0,7	0,9	0,8	0,7	0,7	0,5	0,6	0,6	0,7	0,7
7	0,4	0,3	0,3	0,5	0,5	0,5	0,4	0,4	0,4	0,4	0,6	0,6
8	0,4	0,3	0,3	0,3	0,4	0,4	0,4	0,5	0,8	0,9	0,9	0,8
9	0,3	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,6	0,7	1,0	0,9
10	0,3	0,5	0,5	0,4	0,4	0,4	0,5	0,5	0,6	0,5	0,5	0,6
11	0,5	0,5	0,6	0,6	0,7	0,7	0,8	0,8	1,1	0,7	1,0	0,8
12	0,3	0,4	0,5	0,5	0,5	0,6	0,6	0,5	0,5	0,5	0,5	0,7
13	0,4	0,3	0,3	0,3	0,3	0,5	0,3	0,4	0,6	0,7	0,7	0,7
14	0,3	0,8	0,6	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,5	0,6
15	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,4	0,4	0,5	0,7
16	0,3	0,4	0,3	0,3	0,3	0,3	0,3	0,4	0,4	0,6	0,5	0,7

On considère maintenant les données comme provenant d'un seul parc et avec des traitements ayant tous démarrés au même moment dans le temps. La figure 12 est une représentation des données pour quatre porcs choisis aléatoirement.

À partir de ces mêmes données, il est possible d'analyser le cumul de la consommation quotidienne. La figure 13 illustre l'exemple des mêmes quatre porcs apparaissant déjà dans la figure 12.

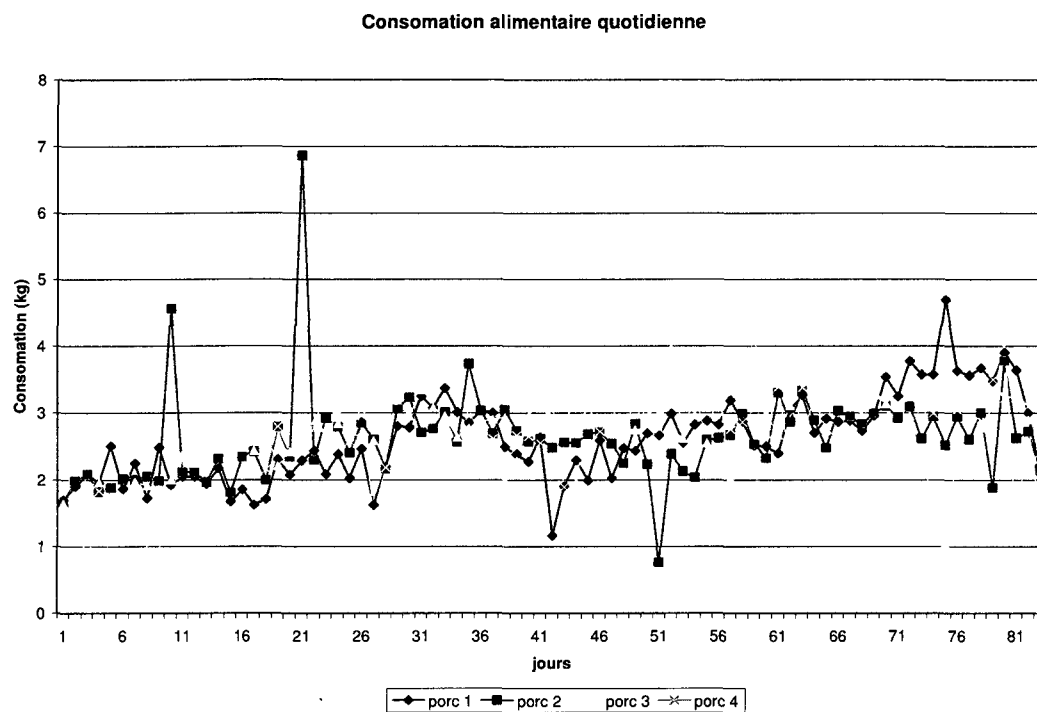


Figure 12: Consommation alimentaire quotidienne (en kg) pour quatre porcs

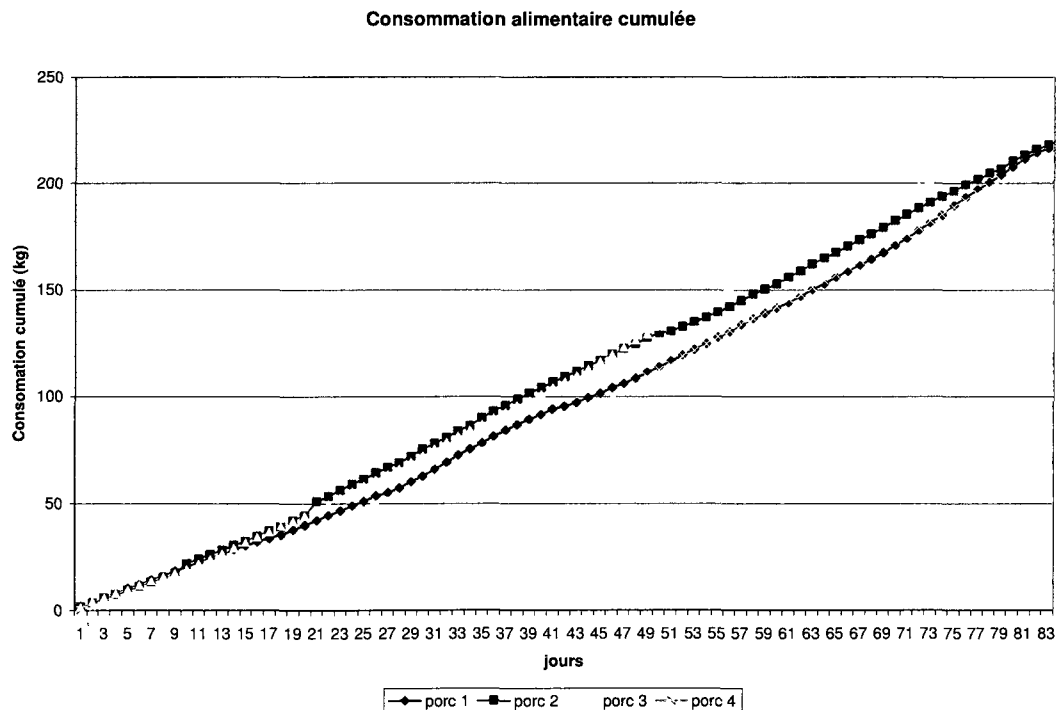


Figure 13: consommation alimentaire cumulée (en kg) pour quatre porcs

Les deux figures précédentes démontrent que, pour le même niveau d'information, les données cumulées sont plus lisses et donc plus faciles à estimer. De plus, on a vu au chapitre précédent que l'analyse en composantes principales et l'analyse discriminante de données fonctionnelles sont facilitées sous l'hypothèse de la différentiabilité des courbes; donc les analyses vont être effectuées à partir des données cumulées.

On a également besoin d'une variable qui indique quand le porc a fait de la fièvre. On utilisera cette variable comme variable approximante de la réelle variable malade ou non malade lors de l'analyse discriminante fonctionnelle. Par contre, comme cela n'est pas un réel test diagnostique pour une maladie en particulier et encore moins une preuve *sine qua non* qu'il y a réellement une maladie, il faudra donc être souple dans la définition de ce que l'on appelle un porc malade. Le choix du critère à utiliser pour considérer un porc malade sera effectué dans la sous-section 4.7.1.

Le tableau 16 présente le nombre de porcs malades ainsi que la proportion qu'ils représentent sur l'ensemble de l'échantillon pour une série de critères potentiels allant d'au moins un jour avec de la fièvre mesurée à au moins sept jours consécutifs avec présence de fièvre.

Tableau 16: Nombre de porcs malades et pourcentages selon différents critères

Nombre de jours consécutifs avec de la fièvre	Nombre de porcs répertoriés répondant au critère	Pourcentage de l'échantillon
Au moins 1 jours (1+)	61	34.5%
Au moins 2 jours (2+)	26	14.7%
Au moins 3 jours (3+)	19	10.7%
Au moins 4 jours (4+)	16	9.0%
Au moins 5 jours (5+)	11	6.2%
Au moins 6 jours (6+)	10	5.6%
Au moins 7 jours (7+)	7	4.0%

4.4 Transformation des variables

Une fois la décision d'effectuer les analyses sur la consommation cumulée, il faut maintenant transformer ces données en données fonctionnelles. Pour ce faire, les fonctions B-splines seront utilisées. Cette façon de faire est inspirée de [Ramsay *et al.* 2002] où l'on retrouve un exemple d'analyses fonctionnelles appliqué sur des données de la croissance humaine. Les données utilisées dans cette analyse sont représentées à la figure 14.

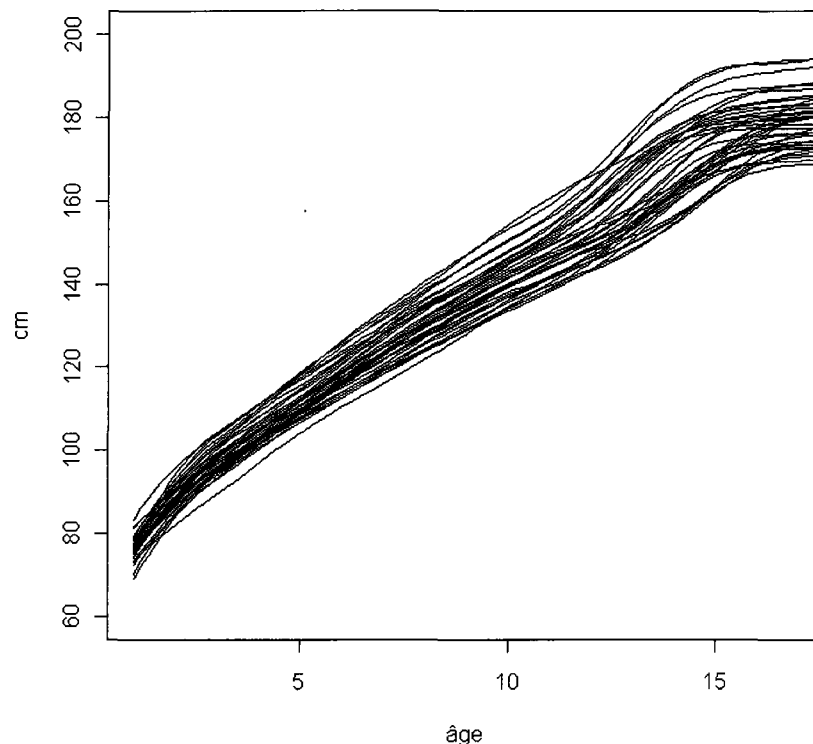


Figure 14: Graphique tiré de l'exemple Ramsay sur des données de croissance humaine

On peut constater que, malgré leur nature différente, les deux jeux de données renferment plusieurs similitudes comme le fait d'être monotone croissant et d'être relativement lisse. Donc, cela nous permet de croire que la méthode des B-splines est toute indiquée pour notre exemple. De plus, les fonctions B-splines ont déjà été utilisées dans un autre cas où l'analyse portait sur des données

alimentaire porcines [Chenguiti 2006]. La prochaine section est dédiée à la définition de ces fonctions, mais tout d'abord, on doit présenter le modèle utilisé.

Soit une suite $t_j \in [a, b] \in \mathbb{R}$, $j : 1, 2, \dots, p$, tel que $a = t_1 < t_2 < \dots < t_p = b$ et n fonctions $x_i(t)$ représentant les observations d'une variable aléatoire fonctionnelles $X(t)$ aux instants t_j . Soit $x_i(t_j) = X(t_j) + \varepsilon_i(t_j)$ pour $i : 1, 2, \dots, n$ et $\varepsilon_i(t_j)$ l'erreur, causée par la modélisation. Pour modéliser $x_i(t)$, on cherche une fonction $f(t)$, à valeur réelle, de classe C^m sur l'intervalle $[a, b]$ tel que $\frac{d^m}{dt^m} f(t) \in L^2$ et qui minimise l'expression, L , suivante :

$$L = \left(\sum_{j=1}^p (f(t_j) - x_i(t_j)) \right)^2 + \lambda \int_a^b \left(\frac{d^m}{dt^m} f(t) \right)^2 dt$$

Le premier terme de cette équation minimise le carré de la distance entre la courbe de lissage et les observations. Le second terme représente la régularité désirée de la courbe de lissage $f(t)$. Il assure aussi que les courbes de lissage sont des éléments qui appartiennent à un espace de Sobolev (voir section 2.6) et que les résultats présentés au chapitre précédent sont valides. Le paramètre λ a pour effet de faire varier l'importance accordée entre l'erreur quadratique et la régularité de la courbe. Si λ tend vers zéro l'équation revient à minimiser l'erreur quadratique et si λ tend vers l'infini, elle revient à négliger totalement la qualité de l'approximation au profit de la régularité de la courbe.

4.4.1 Les fonctions B-splines

Soit une suite de nombre réels t_j avec $j : 1, 2, \dots, p$ tel que $a = t_1 < t_2 < \dots < t_p = b$.
Chacun des t_j est appelé un nœud. Soit aussi $\omega_{i,j}(t)$ des fonctions définies comme suit :

$$\omega_{i,j}(t) = \frac{t - t_i}{t_{i+j} - t_i} \text{ pour tout } t \in [a, b] \text{ et } i, j \text{ deux entiers tels que } j = 1, 2, \dots, p - i$$

On définit par récurrence sur k les fonctions B-splines $B_{i,j}$ pour $i = 1, 2, \dots, n - k - 1$ par la relation suivante :

$$B_{i,1}(t) = \begin{cases} 1 & \text{si } t \in [t_i, t_{i+1}[\\ 0 & \text{si } t \notin [t_i, t_{i+1}[\end{cases}$$

Donc, pour $k \geq 1$ on a la récurrence suivante :

$$B_{i,k}(t) = \omega_{i,k}(t)B_{i,k-1}(t) + (1 - \omega_{i+1,k}(t))B_{i+1,k-1}(t)$$

On note que les fonctions B-splines, $B_{i,j}$, sont des fonctions polynômiales de degré inférieur ou égale à k sur chacun des intervalles $[t_i, t_{i+1}[$. De plus elles appartiennent à la classe des fonctions C^k au voisinage de chaque nœud.

Les fonctions B-splines constituent une base (parmi d'autres) de l'espace vectoriel des fonctions définies sur l'intervalle $[t_1, t_{n-k}]$, polynômiales de degré inférieur ou égal à k sur chaque intervalle $[t_i, t_{i+1}[$ [Risler 1991].

4.4.2 Principales propriétés des fonctions B-splines

On trouvera ci-dessous les principales propriétés qui caractérisent les fonctions B-splines :

1. La fonction $B_{j,k}$ est, sur chaque intervalle $[t_i, t_{i+1}[$, un polynôme de degré k .

Support :

2. la fonction $B_{i,k}$ s'annule en dehors de l'intervalle $[t_i, t_{i+k}[$;
3. la fonction $B_{i,k}$ s'annule aussi en t_i sauf si $t_i = t_{i+1} = \dots = t_{i+k} = t_{i+k+1}$ auquel cas $B_{i,k}(t_i) = 1$;
4. $0 < B_{i,k}(t) \leq 1$ pour tout $t \in]t_i, t_{i+k}[$;
5. sur l'intervalle $]t_i, t_{i+k}[$, la fonction $B_{i,k}$ ne prend la valeur 1 que si $t_{i+1} = \dots = t_{i+k}$ et en ce point seulement.

Note : le cas où $B_{i,k}(t_i) = 1$ ne s'applique pas dans notre exemple car $a = t_1 < t_2 < \dots < t_p = b$.

Partition de l'unité :

6. sur l'intervalle $[t_k, t_{n-k}[$ on a : $\sum_{i=0}^{p-k-1} B_{i,k} = 1$

Différentiabilité :

7. la fonction $B_{i,k}$ est C^∞ à droite de chaque point ;

8. au voisinage d'un nœud de multiplicité r ; la fonction $B_{i,k}$ est seulement de classe C^{k-r} .

On remarque que la proposition de support découle directement de la définition des fonctions B-splines. Par contre les propositions à propos de la partition de l'unité et de la différentiabilité ne sont pas triviales et nécessite une démonstration.

Preuve par récurrence de la partition de l'unité : Elle est clairement vraie pour $k = 0$. On la suppose établie pour les B-splines de degré inférieur ou égale à $k - 1$. Cela permet d'écrire :

$$\sum_{i=0}^{p-k-1} B_{i,k} = \omega_{0,k} B_{0,k-1} + (1 - \omega_{1,k}) B_{1,k-1} + \omega_{1,k} B_{1,k-1} + (1 - \omega_{2,k}) B_{2,k-1} + \dots + \omega_{p-k-1,k} B_{p-k-1,k-1} + (1 - \omega_{p-k,k}) B_{p-k,k-1}$$

$$\sum_{i=0}^{p-k-1} B_{i,k} = (-1 + \omega_{0,k}) B_{0,k-1} + \sum_{i=0}^{p-k} B_{i,k-1} (-1 + \omega_{p-k,k}) B_{p-k,k-1}.$$

D'après l'hypothèse de récurrence, la somme $\sum_{i=0}^{p-k} B_{i,k-1} = 1$ sur l'intervalle

$[t_{k-1}, t_{p-k+1}[$, donc il en découle qu'il vaut aussi 1 sur l'intervalle $[t_k, t_{n-k}[$.

D'après la propriété du support, la fonction $B_{0,k-1}$ est nulle hors de l'intervalle

$[t_0, t_k[$, donc la somme $\sum_{i=0}^{p-k-1} B_{i,k} = 1$ sur l'intervalle $[t_k, t_{n-k}[$.

Pour démontrer la propriété de différentiabilité, on doit tout d'abord introduire deux lemmes essentiels démontrés dans [Risler 1991]. Pour la suite, si une fonction f définie sur les réels admet des limites à droite et à gauche, on les notera $f(t+)$ et $f(t-)$.

Lemme : Si $t_{i-1} < t_i = t_{i+1} = \dots = t_{i+r-1} < t_{i+r}$, alors $B_{i-1,r}(t_{\pm}) = 1$. Par conséquent, toutes les fonctions $B_{i,k}$ pour $k \geq r$ sont continues en t_j .

Lemme : Pour tout $k \geq 0$, la fonction $B_{i,k}$ est dérivable à droite et la dérivé est égale à :

$$B'_{i,k} = k(\omega'_{i,k} B_{i,k-1} - \omega'_{i+1,k} B_{i+1,k-1})$$

que l'on peut écrire :

$$B'_{i,k}(t) = k \left(\frac{B_{i,k-1}(t)}{t_{i+k} - t_i} - \frac{B_{i+1,k-1}(t)}{t_{i+k+1} - t_{i+1}} \right)$$

Donc, ce dernier lemme montre l'existence de dérivées à droite de tous ordres.

Maintenant soit t un nœud de multiplicité r , alors les B-splines de degré r sont continues en t . Toujours avec le lemme précédent, on peut affirmer que les B-splines de degré $r+1$ sont de classe C^1 au voisinage de t . Par récurrence, on montre ainsi que les B-splines de degré $k \geq r$ sont de classe C^{k-r} au voisinage d'un nœud de multiplicité r .

4.4.3 Approximation par des fonctions B-splines

On définit une approximation B-spline de degré k par toute fonction qui pour tout $t \in [a, b]$ fait correspondre :

$$x_k(t) = \sum_i c_i B_{i,k}(t)$$

où $c_i, i \in N$ est une suite de nombre réels.

Par la section précédente, l'approximation par des fonctions B-splines possède les propriétés suivantes :

1. Les composantes $x_k(t)$ sont des polynômes de degré k sur chacun des intervalles $[t_i, t_{i+2}]$;
2. En chaque point $t_i \in [a, b]$, $x_k(t)$ est de classe C^k .

4.4.4 Estimation

Une fois la base choisie, on cherche à estimer les coefficients de l'approximation qui minimise l'expression L . En disposant de n fonctions $x_i(t)$ et pour un nombre K de termes d'approximations, on exprime la solution cherchée sous la forme suivante :

$$y(t) = \sum_{k=1}^K c_k \phi_k(t)$$

On pose $c = (c_1, c_2, \dots, c_K)$ et $\varphi = (\phi_1(t), \phi_2(t), \dots, \phi_K(t))^T$, ce qui permet d'écrire $y(t)$ comme suit :

$$y(t) = c^T \varphi = c \varphi^T$$

On pose aussi la matrice Φ de terme général $\phi_k(t_j)$, avec $k : 1, 2, \dots, K$ et $j : 1, 2, \dots, n$, tel que :

$$\Phi = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_K(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_K(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_n) & \phi_2(t_n) & \cdots & \phi_K(t_n) \end{pmatrix}$$

On note aussi $X_i = (x_i(t_1), x_i(t_2), \dots, x_i(t_n))$ et W la matrice diagonale des poids de terme général w_i affectant chacune des observations $x_i(t_j)$ avec $i : 1, 2, \dots, n$ et $j : 1, 2, \dots, p$. L'expression L pour l'ensemble des observations $x_i(t)$ devient :

$$L = \sum_{i=1}^n w_i (x_i(t_j) - y(t_j))^2 + \lambda \int_a^b \left(\frac{d^m}{dt^m} y(t) \right)^2 dt$$

que l'on peut écrire avec la notation introduite précédemment :

$$L = (X - \Phi c)^T W (X - \Phi c) + \lambda \int_a^b \left(\frac{d^m}{dt^m} c^T \varphi \right)^2 dt$$

On peut écrire cette dernière intégrale comme :

$$\int_a^b \left(\frac{d^m}{dt^m} c^T \varphi \right)^2 dt = \int_a^b \left(c^T \frac{d^m}{dt^m} \varphi \frac{d^m}{dt^m} \varphi^T c \right) dt = c^T \left[\int_a^b \left(\frac{d^m}{dt^m} \varphi \frac{d^m}{dt^m} \varphi^T \right) dt \right] c$$

Si on pose :

$$R = \int_a^b \left(\frac{d^m}{dt^m} \varphi \frac{d^m}{dt^m} \varphi^T \right) dt$$

on peut écrire l'expression L comme :

$$L = (X - \Phi c)^T W (X - \Phi c) + \lambda c^T R c$$

À partir de cette dernière expression que l'on dérive par rapport à c et que l'on pose égale à zéro, il est plus facile d'isoler c :

$$\frac{d}{dc} L = \frac{d}{dc} \left[(X - \Phi c)^T W (X - \Phi c) + \lambda c^T R c \right] = 0$$

$$-2\Phi^T W X + 2\Phi^T W \Phi c + 2\lambda R c = 0$$

$$-\Phi^T W X + \Phi^T W \Phi c + \lambda R c = 0$$

$$\Phi^T W X = \Phi^T W \Phi c + \lambda R c$$

$$\Phi^T W X = (\Phi^T W \Phi + \lambda R) c$$

ce qui implique que :

$$\hat{c} = (\Phi^T W \Phi + \lambda R)^{-1} \Phi^T W X$$

On pose finalement $S_{\phi, \lambda}$, la matrice, nommée matrice de lissage, telle que :

$$S_{\phi, \lambda} = (\Phi^T W \Phi + \lambda R)^{-1} \Phi^T W$$

donc, la solution matricielle peut s'écrire comme :

$$\hat{Y} = S_{\phi, \lambda} X$$

4.5 Lissage des données

Dans cette section, on transforme les données de consommation cumulées en données fonctionnelles. Les analyses qui vont suivre ont été effectuées à l'aide du logiciel R version 2.11.1³. De plus, des textes concernant la programmation et l'utilisation du logiciel R avec des données fonctionnelles tels [Ramsay 2009] ou [Clarkson 2005] ont été consultés.

On utilise les fonctions B-splines d'ordre six comme base et on impose une contrainte de régularité sur la dérivée d'ordre quatre. Cela permet d'assurer l'existence et la régularité des dérivées d'ordres inférieures. Donc, avec ces paramètres, on s'assure de pouvoir travailler avec des courbes de vitesse et d'accélération [Ramsay *et al.* 2002]. On rappelle que l'on dispose de données pour $n=177$ porcs et cela pendant 83 jours, donc l'expression L à minimiser, subjacente au modèle, s'écrit comme suit :

$$L = \sum_{i=1}^{177} w_i (x_i(t) - y_i(t))^2 + \lambda \int_1^{83} \left(\frac{d^4}{dt^4} y(t) \right)^2 dt$$

où $w_i = \frac{1}{177}$, ce qui signifie que chaque porc est affecté du même poids.

Comme $y(t) = c^T \phi$, ceci revient à trouver les fonctions $\phi_k(t)$ qui minimisent l'erreur quadratique moyenne et qui tiennent compte de la contrainte de régularité. Il faut aussi déterminer le λ qui optimise la qualité de l'approximation tout en tenant compte de la contrainte de régularité. Dans cet exemple, on utilise un critère basé sur la validation croisée généralisée, VCG, [Craven *et al.*, 1979] qui est définie comme suit :

³ Disponible à l'adresse suivante : <http://www.r-project.org/index.html>

$$VCG(\lambda) = \frac{n^{-1}(X - \Phi c)'W(X - \Phi c)}{\left(n^{-1}tr(I - S_{\Phi, \lambda})\right)^2}$$

où

$$X = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \dots & x_{177}(t_1) \\ x_1(t_2) & x_2(t_2) & \dots & x_{177}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_{177}(t_{83}) & x_{177}(t_{83}) & \dots & x_{177}(t_{83}) \end{pmatrix}$$

et Φ , c , $S_{\Phi, \lambda}$ et W sont définis comme à la section précédente. Finalement I est la matrice identité 83 par 83.

Comme le numérateur de l'expression $VCG(\lambda)$ représente l'erreur quadratique et que le dénominateur de cette même expression représente la contrainte de régularité, la validation croisée généralisée nous permet de visualiser la valeur de λ qui offre le meilleur compromis entre ces deux éléments. La figure 15 présente les valeurs prises par $VCG(\lambda)$ pour différentes valeurs de λ . On remarque de $\log(\lambda) = -1$ est la valeur où $VCG(\lambda)$ est minimum ce qui implique que l'on utilisera un $\lambda = 10^{-1}$ pour le lissage final des données.

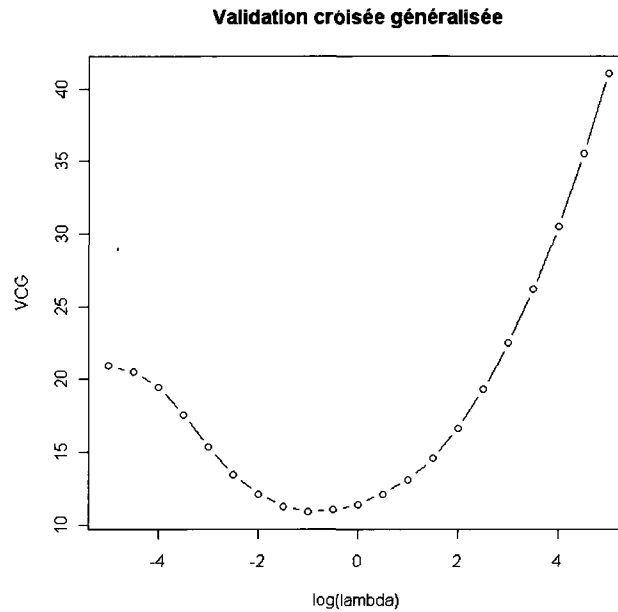


Figure 15: Valeur de la validation croisée généralisée pour différence valeurs de lambda

La figure 16 représente les 177 courbes de consommations cumulées estimées avec le lambda optimal. On s'assure que les observations sont bien estimées en calculant l'erreur quadratique qui est en moyenne de 0,018 et une variance de 0,0002; ce qui est plus qu'acceptable. La figure 17 présente l'erreur quadratique moyenne pour chacun des porcs.

Fonctions estimées pour chacun des porcs

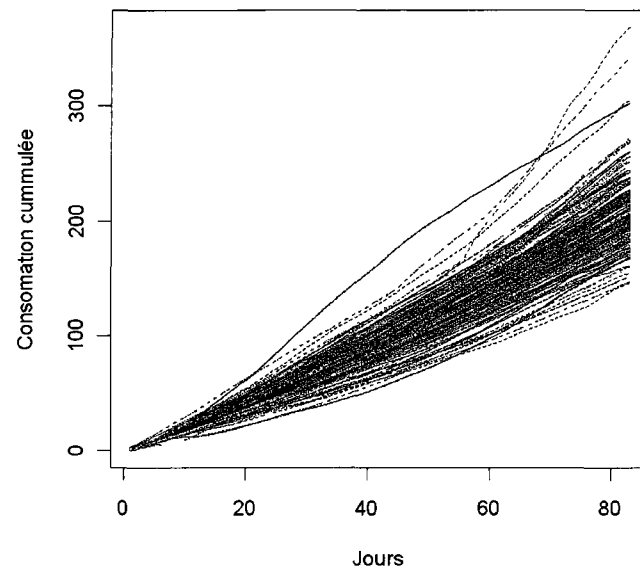


Figure 16: Les courbes de consommation cumulée estimées pour les 177 porcs

**Écart quadratique moyen (EQM)
pour chacun des porcs**

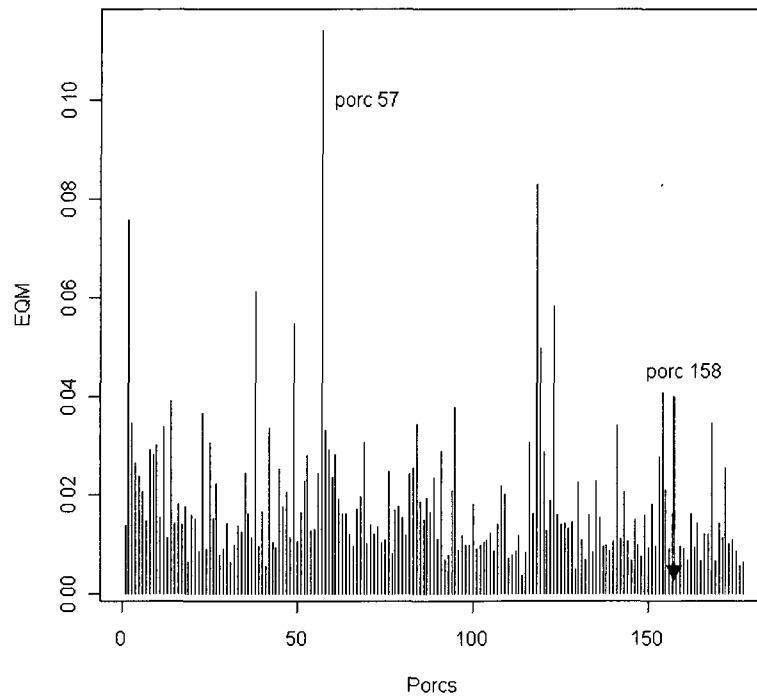


Figure 17: Erreur quadratique moyenne pour chacun des 177 porcs

À partir de cette dernière figure, on remarque que le porc 57 est celui avec l'EQM le plus élevé avec une valeur de 0,11 et que le porc 158 est celui avec la valeur minimum pour ce même indicateur, soit 0,0032. Pour ces deux observations, on peut regarder plus attentivement, à la figure 18, ce que cela signifie graphiquement en termes de l'approximation et des résidus.

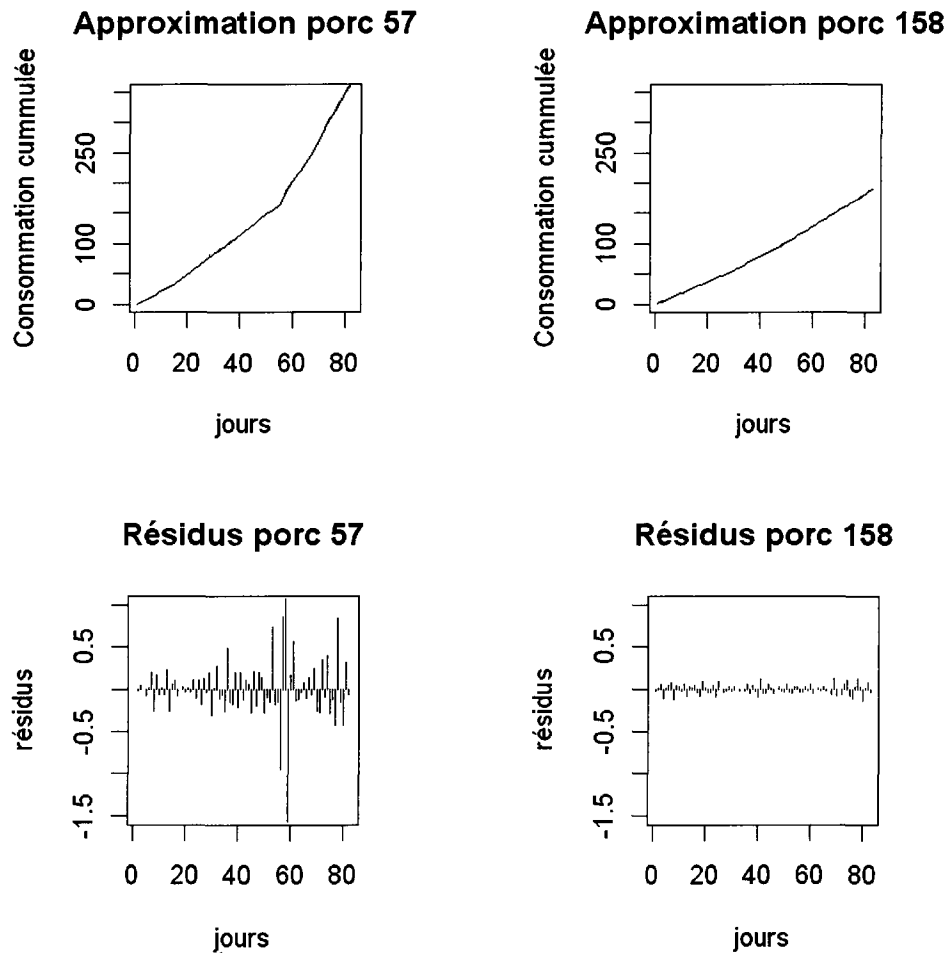


Figure 18: Approximation et résidus pour les deux porcs ayant un EQM extrême

De cette figure, on remarque que, malgré la différence marquée entre les résidus des deux porcs, il est pratiquement impossible de distinguer la courbe de la consommation estimée, en noir, et les points observés, en rouge. Cela est attribué à la différence entre l'échelle de la mesure de la consommation, de 0 à approximativement 350 et celle des résidus de -1,5 à 1. Pour pallier à cet inconvénient, on peut se concentrer sur une plus petite zone du graphique comme le fait la figure 19 pour le porc 57.

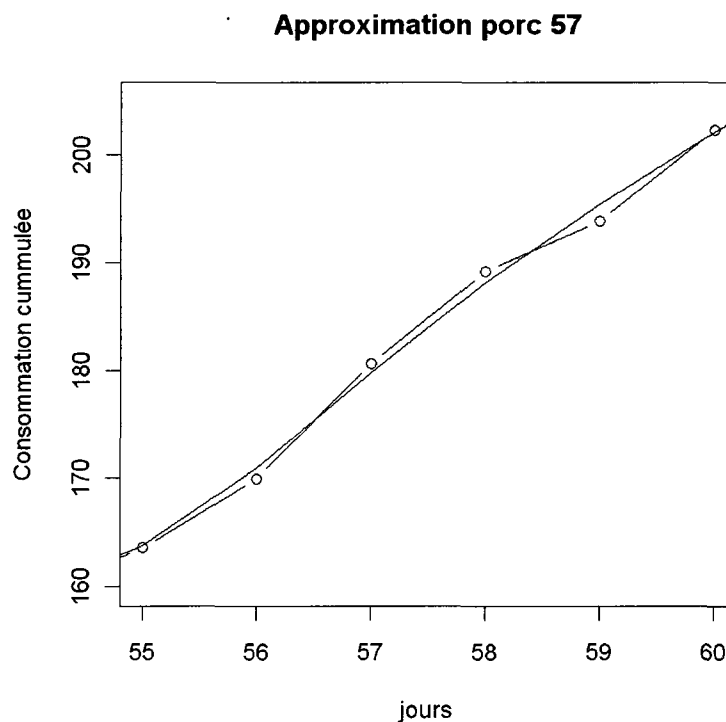


Figure 19: Agrandissement du graphique de la fonction d'approximation pour un porc

La figure 19 est un agrandissement de la zone où les résidus sont les plus grand pour le porc avec l'erreur quadratique moyenne la plus élevée et même dans ce cas, on observe que la qualité de l'ajustement est bonne. Donc, on en déduit que l'approximation, avec comme base les B-splines et un paramètre $\lambda = 10^{-1}$ est convenable et que l'on peut poursuivre avec l'analyse en composantes principales fonctionnelles.

4.6 Application de l'analyse en composantes principales fonctionnelles (ACPF) aux données de consommation cumulée porcine et à ses dérivées d'ordre un et deux

Dans cette section, on présente les résultats de l'analyse en composantes principales fonctionnelle des données de consommation cumulée porcine. De plus, comme on considère des fonctions, on effectuera aussi les mêmes analyses pour les première et deuxième dérivées, soient les courbes représentant la vitesse et l'accélération de la consommation cumulée.

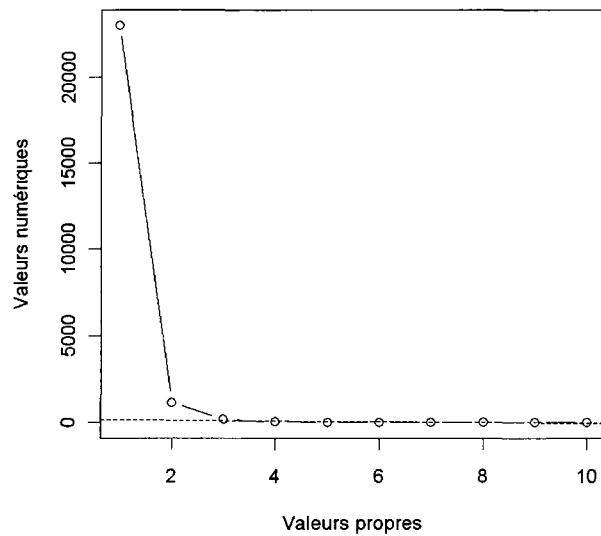
Ces trois analyses sont effectuées pour pouvoir représenter les observations dans un sous-espace où le pourcentage de la variabilité des données expliquée est optimal, par les différentes composantes principales fonctionnelles, ce qui permet de mieux comprendre ce qui influence les observations.

4.6.1 ACPF des courbes de consommation cumulée

On doit d'abord déterminer combien de composantes principales sont nécessaires pour prendre en compte la variabilité des observations. Pour ce faire, on regarde les valeurs propres de MV tel que présentée dans la figure 20.

En observant la figure 20, on remarque que seul les deux premières valeurs propres, et particulièrement la première, se démarquent des autres, car dès la troisième elles sont pratiquement égales entre elles et aussi proches de zéro. Cela signifie que seulement les deux premières composantes principales seront retenues pour l'analyse, car elles comprennent pratiquement la totalité de la variabilité des observations, soit 99%.

**Les 10 premières valeurs propres associées à ACPF
des courbes de la consommation cumulée**



**Figure 20: Les 10 premières valeurs propres de l'analyse en composantes principales
fonctionnelles de la consommation cumulée**

On peut maintenant effectuer l'analyse en composantes principales. La figure 21 présente les deux premières composantes principales qui sont les fonctions propres associées aux deux premières valeurs propres.

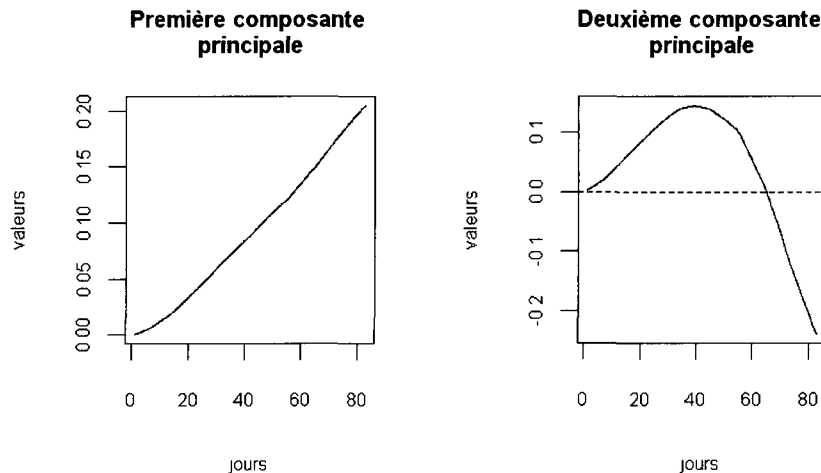


Figure 21: Première et de la deuxième composantes principales

La figure 22 présente l'effet de la première composante principale comme une perturbation positive, « + », et une perturbation négative, « - », de deux écarts-types, noté σ , autour de la moyenne des courbes du cumul de la consommation alimentaire ($\bar{X} \pm 2\sigma u_1$).

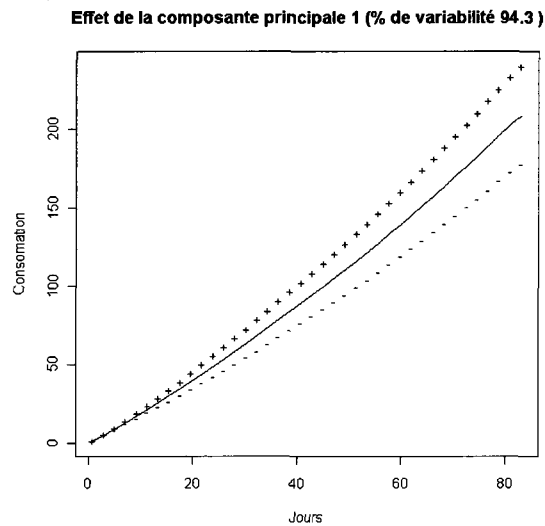


Figure 22: Effet de la première composante principale sur la moyenne de la consommation cumulée

La figure 23 est équivalente à la figure 22, mais cette fois pour la deuxième composante principale.

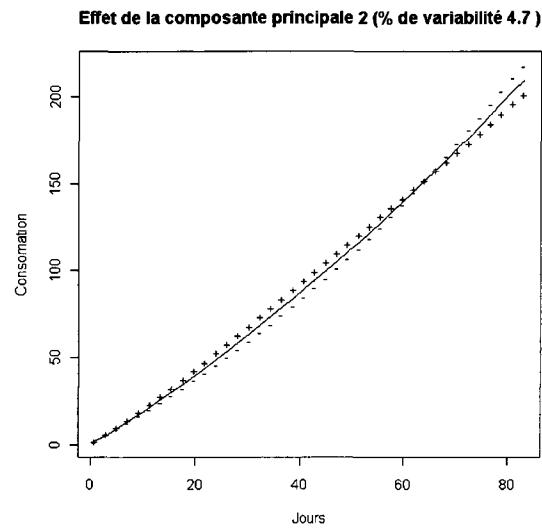


Figure 23 : Effet de la deuxième composante principale sur la moyenne de la consommation cumulée

Ce type de graphiques, telle la figure 22 et 23, permet de bien visualiser l'effet des différentes composantes principales et de mieux comprendre ce qu'elles représentent.

On peut aussi projeter les observations dans le plan constitué par les deux premières composantes principales. Ceci est présenté dans la figure 24.

Projection des observations dans le plan des deux premières composantes principales

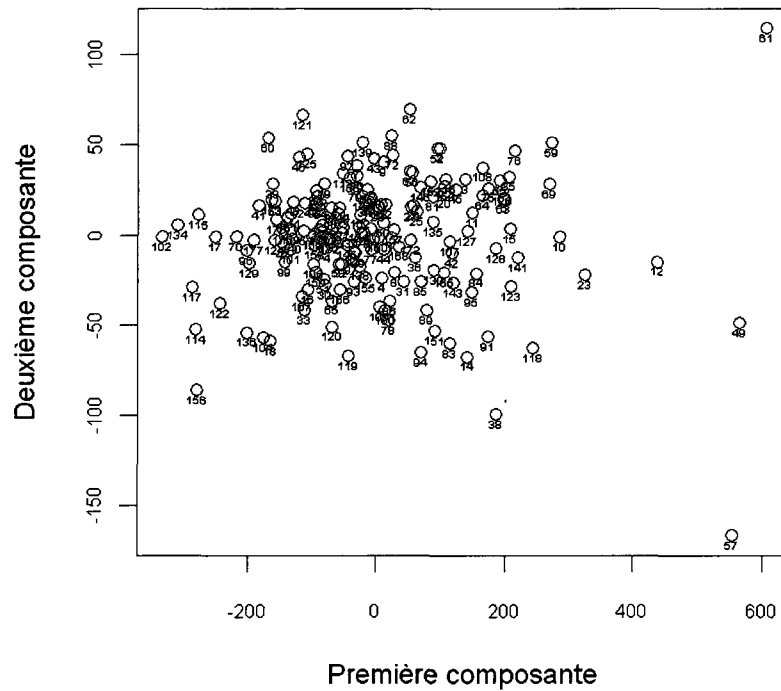


Figure 24: Projection des observations dans le plan constitué par le premier et le deuxième axe principal

On peut aussi agrandir la zone centrale de la figure 24 pour mieux identifier les observations qui sont projetées à l'origine du plan engendré par les deux premières composantes principales. Ceci est fait à la figure 25.

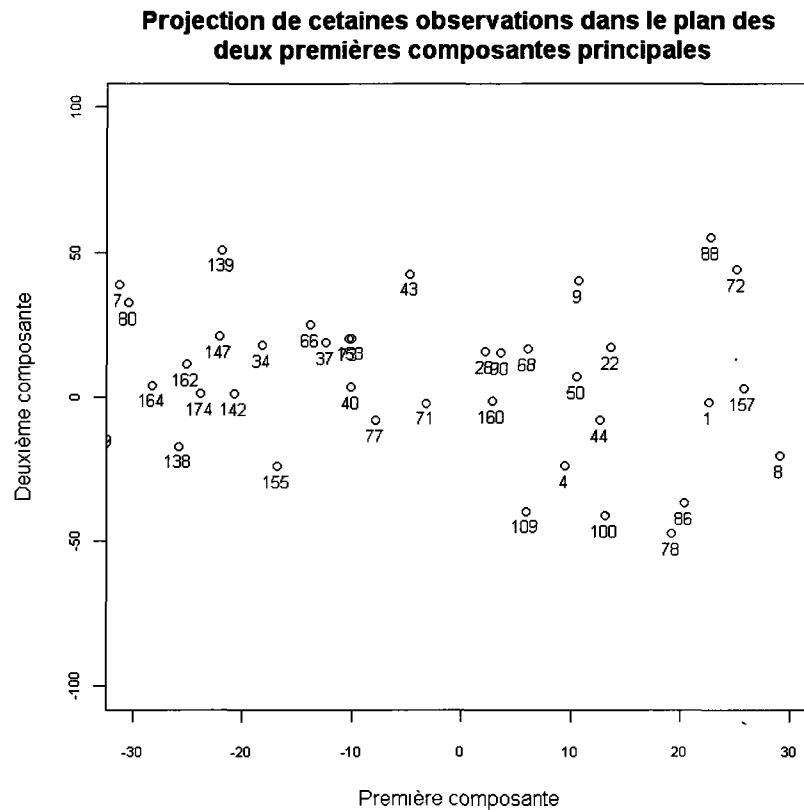


Figure 25 : Projections des Porcs près de l'origine du plan constitué par le premier et le deuxième axe principal

À la lumière des différentes figures présentées, on peut décrire, d'un point de vue statistique et non de celui d'un expert en alimentation porcine, ce que représente chacun des axes principaux.

Le premier axe principal, qui prend en compte 94,3% de la variabilité, représente le niveau de la consommation cumulée total de chacun des porcs tout au long des 83 jours de l'expérience. Pour corroborer ceci, on présente dans la partie gauche de la figure 26 qui représente deux des porcs les mieux discriminés selon la première composante principale, soit le porc numéro 49 (ligne pleine) et le porc numéro 102 (ligne pointillée). De plus, on y retrouve aussi dans la partie droite de cette même figure les porcs numéro 4 (ligne pleine) et 8 (ligne pointillé) qui eux sont très près l'un de l'autre quand ils sont projetés dans le plan des deux premiers axes principaux. Donc, plus la coordonnée sur de la première

composante principale est faible (forte), moins (plus) le porc à consommé de kilogramme de nourriture durant la période d'observation.

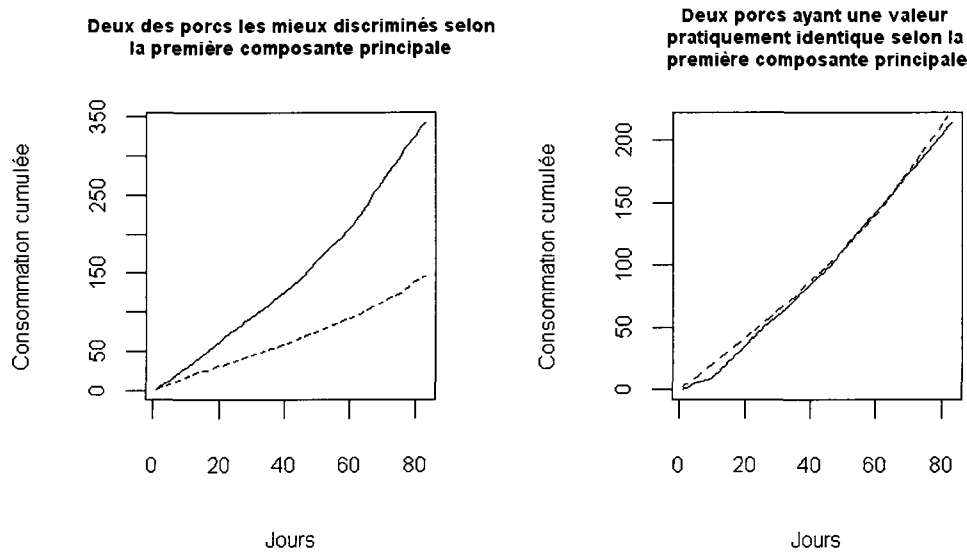


Figure 26: Consommation cumulée de deux porcs bien discriminés et de deux autres ayant des coordonnées proches selon le premier axe principal

Pour ce qui est du deuxième axe principal, il est clair que la différence entre les porcs est moins évidente à faire car il prend en compte seulement de 4,7% de la variabilité totale. Par contre, on observe qu'il représente le changement dans la quantité de nourriture consommée au cours de la période d'observation comme en témoigne la concavité des courbes apparaissant à la figure 27. Cela ce traduit par le fait qu'un porc qui a tendance à consommer plus dans la première partie de l'expérience, soit environ les 40 premiers jours, va généralement diminuer la quantité de nourriture qu'il absorbe pour le reste de la période d'observation. La figure 27, qui représente les porcs numéros 57 et 38 (ligne pleine) qui possèdent des coordonnées fortement négatives par rapport au deuxième axe principal et les porcs numéros 61 et 62 (ligne pointillée) qui eux ont des coordonnées fortement positives par rapport à ce même axe, illustre bien l'explication avancée.

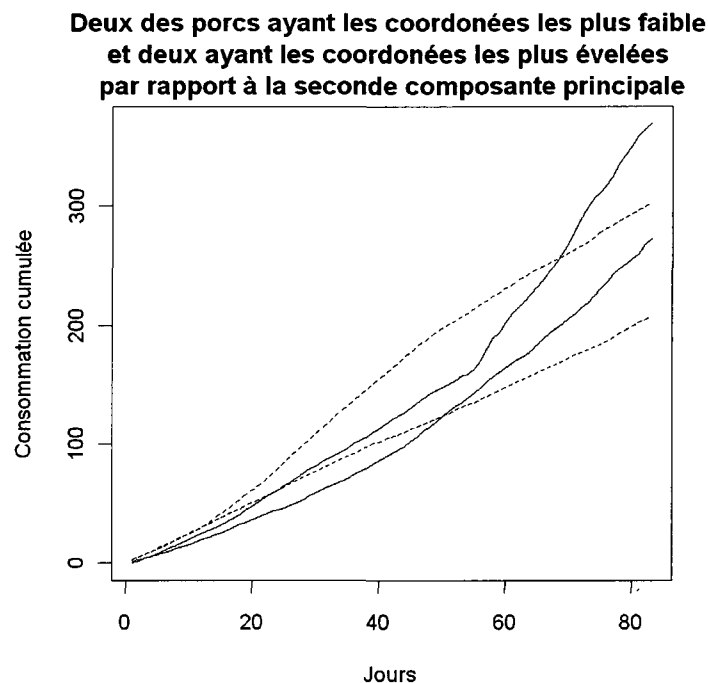


Figure 27: Quatre porcs ayant des coordonnées extrêmes par rapport à la deuxième composante principale

Donc, plus la coordonnée sur de la deuxième composante principale est faible (forte), plus (moins) le porc à consommé de kilogrammes de nourriture durant la première partie de la période d'observation.

4.6.2 ACPF de la première dérivée des courbes de consommation cumulée

On effectue maintenant une analyse en composantes principales sur la dérivée première, qui représente la vitesse à laquelle les fonctions de consommation alimentaire porcine cumulée varient. La figure 28 représente ces dérivées pour les 177 porcs analysés.

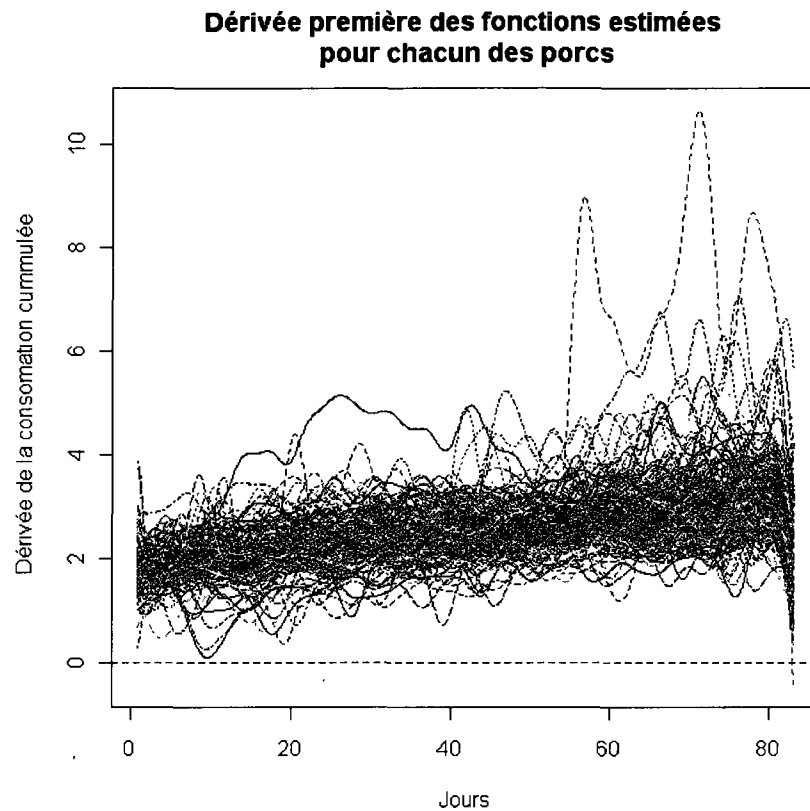


Figure 28: Dérivée première de la consommation alimentaire cumulée pour tous les porcs

Dans cette analyse, on suit les mêmes étapes et dans les mêmes objectifs que l'analyse de la consommation cumulée. Donc, il faut déterminer combien de composantes principales il faut estimer en regardant la valeur des différentes valeurs propres illustrées dans la figure 29.

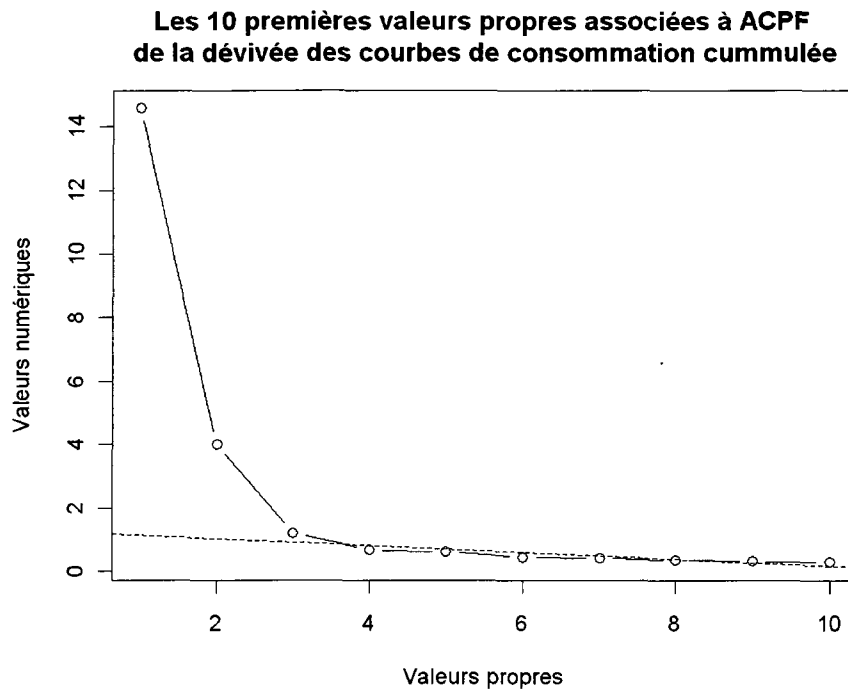


Figure 29 : Les 10 premières valeurs propres de l'analyse en composantes principales fonctionnelles de la première dérivée des courbes de consommation cumulée

Dans ce cas, on retient aussi seulement les deux premières composantes principales pour effectuer l'ACPF. La figure 30 présente ces deux composantes.

Dans cette analyse, les deux premières composantes principales prennent en compte 73% de la variabilité totale des données. Pour mieux comprendre leur signification respective, on présente, à l'aide des figure 31 et 32, l'effet d'une petite variation de deux écarts-types positive (« + ») ou négative (« - ») aux abords de la moyenne et cela pour chacune des deux composantes principales.

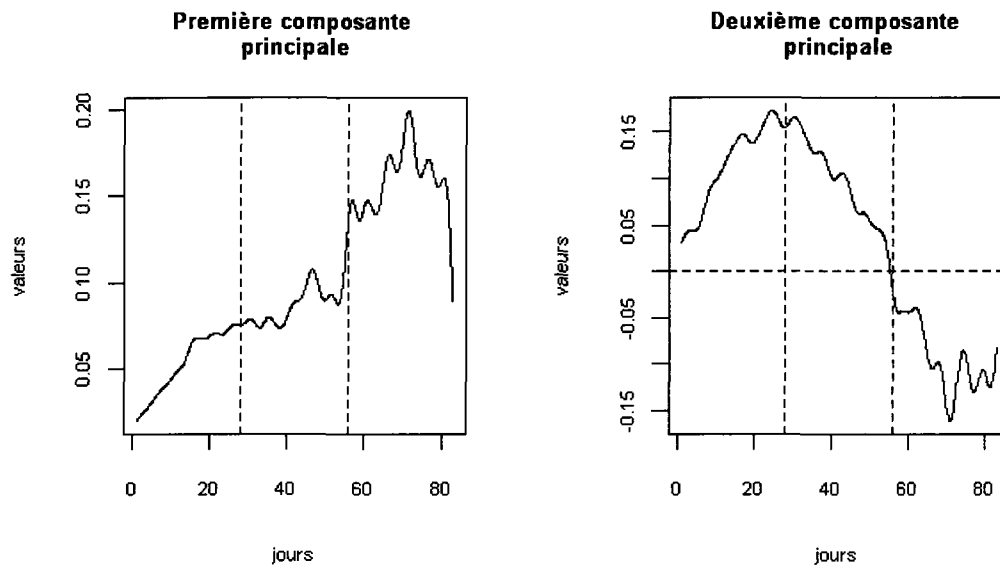


Figure 30: Première et deuxième composantes principales pour L'ACPF de la première dérivée des courbes de consommation cumulée

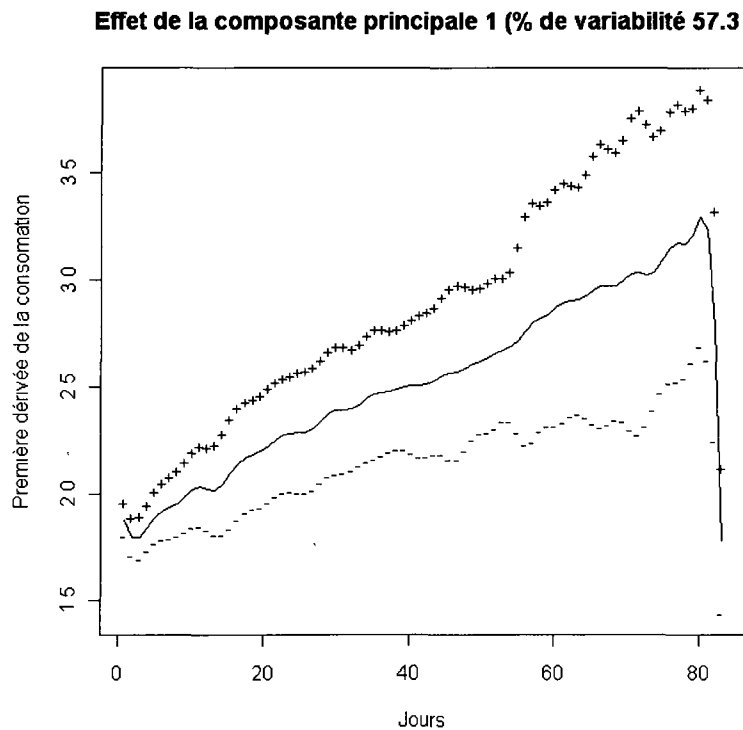


Figure 31: Effet de la première composante principale sur la moyenne

Dans cette analyse, on remarque l'effet des trois différentes phases alimentaires d'une durée de 28 jours chacune. Elles sont délimitées par les lignes pointillées dans les graphiques de la figure 30. Elle est particulièrement apparente pour la première composante principale qui représente 57,3% de la variabilité totale. On voit qu'elle augmente de façon pratiquement constante durant la première phase pour ensuite continuer d'augmenter, mais plus lentement durant la deuxième phase. Pour ce qui est de la troisième phase, on remarque qu'elle reprend de la vitesse jusqu'à atteindre un point, au alentour du jour 80, où elle chute drastiquement. Ce point représente la fin de la période de croissance des porcs, donc un moment où la vitesse de la consommation cumulée a tendance à diminuer rapidement.

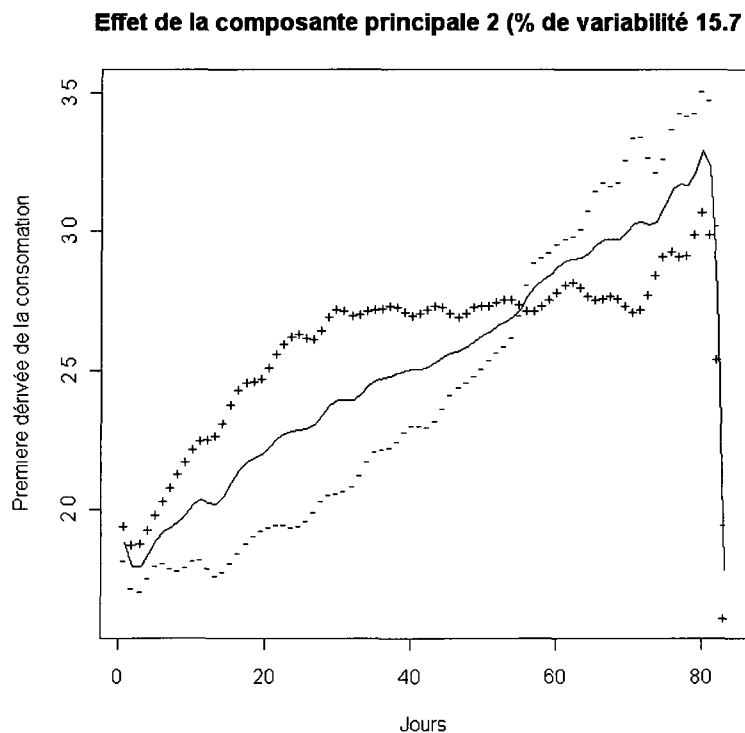


Figure 32 : Effet de la deuxième composante principale sur la moyenne

On peut également projeter les observations dans le plan constitué par les deux premiers axes principaux. Ceci est présenté par la figure 33.



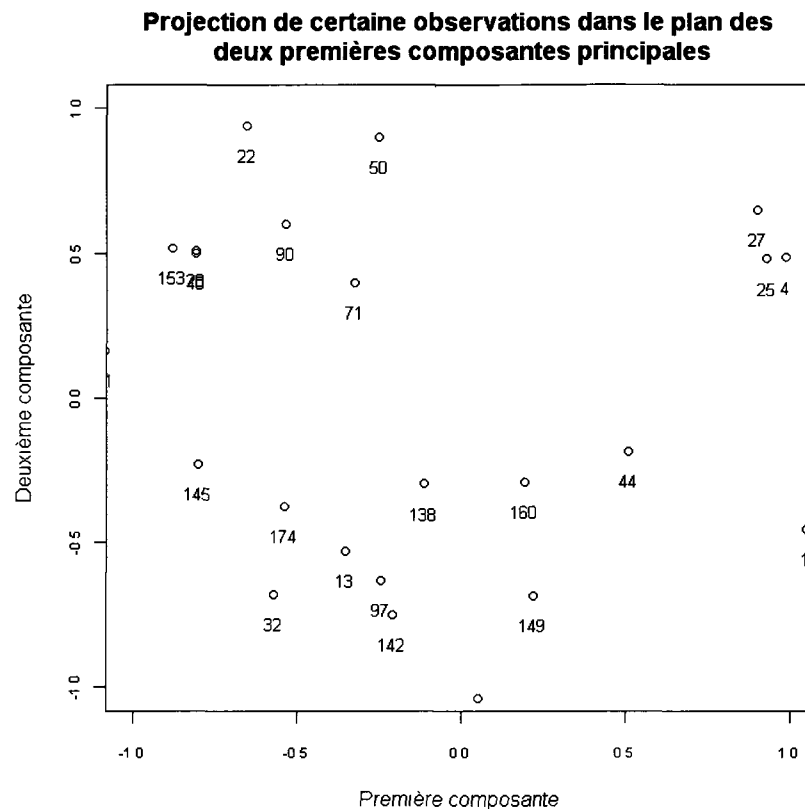


Figure 34 : Projections des observations près de l'origine du plan constitué par le premier et le deuxième axe principal

La figure 34 est un agrandissement de la figure 33 autour de l'origine. Elle permet de différencier les observations qui sont projetées près les unes des autres autour de l'origine du premier plan principal.

La description de cette dernière figure est similaire à son homonyme de l'ACPF de la consommation cumulée. Plus la coordonnée d'un porc, par rapport au premier axe principal, est faible (élevée), moins (plus) sa vitesse de consommation fut élevée durant la période d'observation. Finalement, plus la coordonnée d'un porc, par rapport au deuxième axe principal, est faible (élevée), moins (plus) sa vitesse de consommation fut élevée durant la première période d'observation et le contraire est vrai pour la seconde partie.

4.6.2 ACPF de la deuxième dérivée des courbes de consommation cumulée

La dernière analyse en composantes principales fonctionnelles que l'on a effectué est celle de la dérivée seconde, soit l'accélération, des courbes de consommation cumulée. La figure 35 illustre les fonctions pour l'ensemble des 177 observations que l'on va analyser dans cette section.

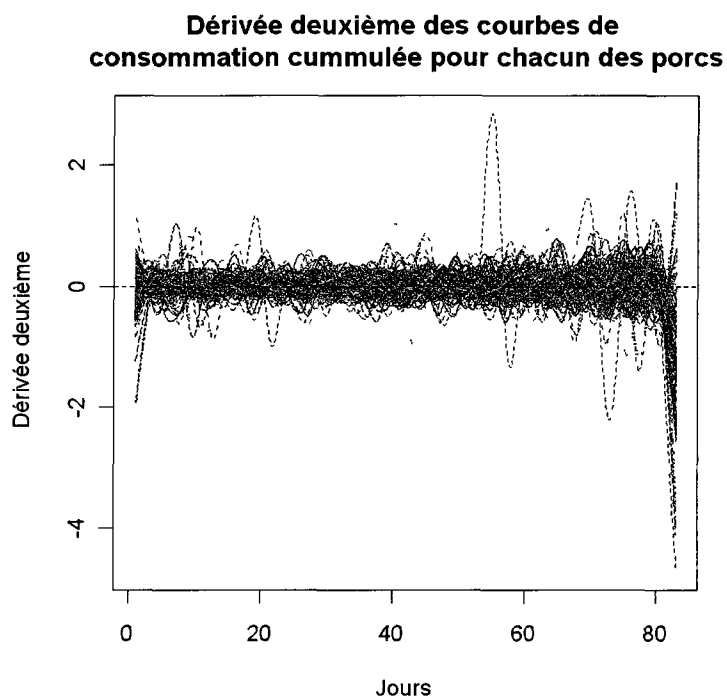


Figure 35: Dérivée seconde de la consommation alimentaire cumulée pour tous les porcs

Quant à elle, la figure 36 présente les 20 premières valeurs propres de la matrice MV associée à l'ACPF de la deuxième dérivée des courbes de consommation cumulée.

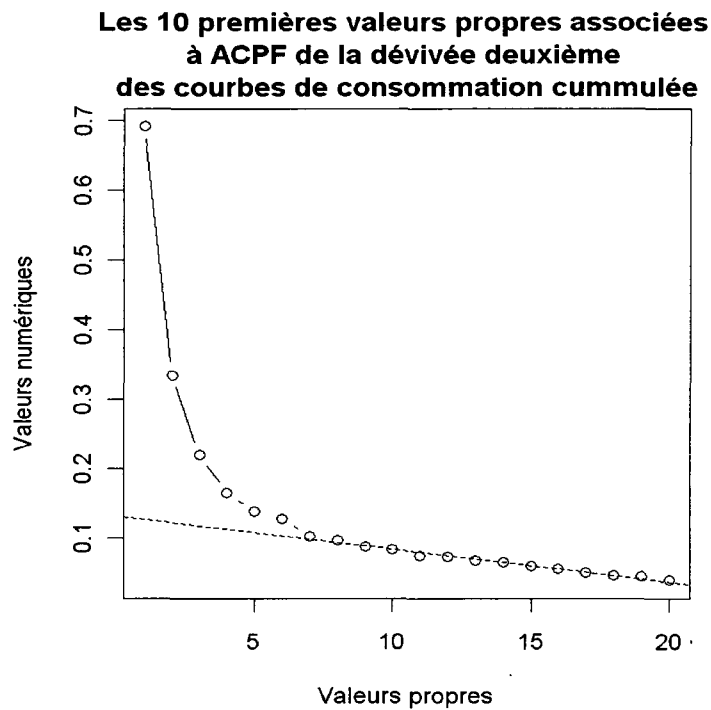


Figure 36 : Les 20 premières valeurs propres de l'analyse en composantes principales de la seconde dérivée des courbes de consommation cumulée

De cette figure, on remarque que l'on doit prendre en compte un plus grand nombre de valeurs propres avant de voir se dégager une tendance de stabilité. Cela signifie que l'on a besoin d'un espace avec un plus grand nombre de dimension pour représenter un pourcentage acceptable de la variabilité contenue dans la deuxième dérivée des courbes de consommation cumulée. La figure 37 illustre le pourcentage de variabilité expliquée par chacune des 20 premières composantes.

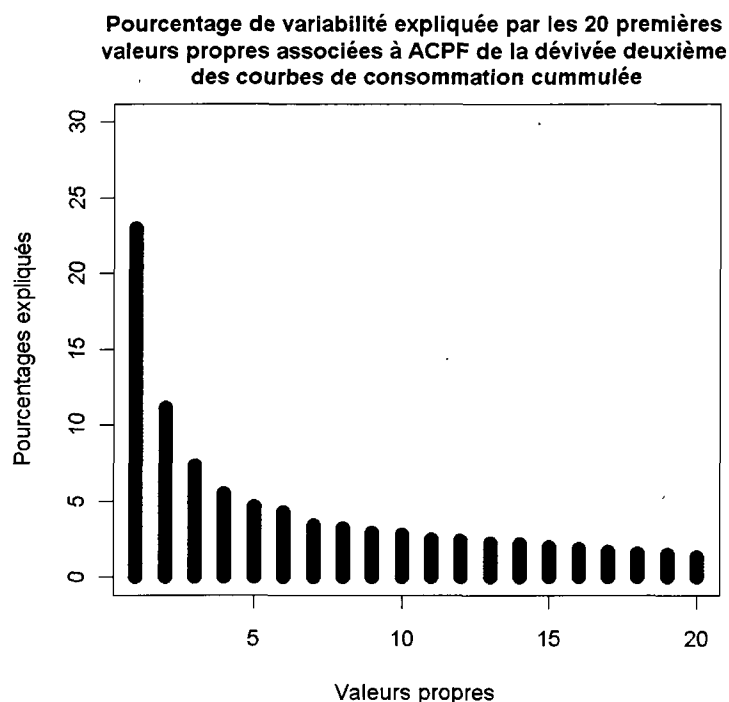


Figure 37 : Pourcentage de variabilité expliquée par les 20 premières valeurs propres

À partir de cette dernière figure, on remarque que la première composante principale prend en compte seulement que 23% de la variabilité comprise dans les courbes. De plus, si on représente les observations dans l'espace engendré par les trois premières composantes principales, il n'y a que 42,4% de la variabilité qui est expliquée et il faut utiliser un espace de plus de cinq dimensions pour avoir un pourcentage de variabilité expliquée supérieur à 50%. Finalement, si on regarde la première composante principale, comme le fait la figure 38, on se rend compte que, par ces nombreuses variations autour de zéro, elle est difficilement interprétable.

Donc, comme l'analyse en composantes principales des courbes de consommation cumulée et de leurs dérivées on déjà fournit des résultats plus probants, on décide de ne pas tenir compte de l'ACPF des secondes dérivées en admettant que cette méthode, telle que décrite précédemment et avec les données que l'on dispose, n'est pas nécessairement la plus appropriée.

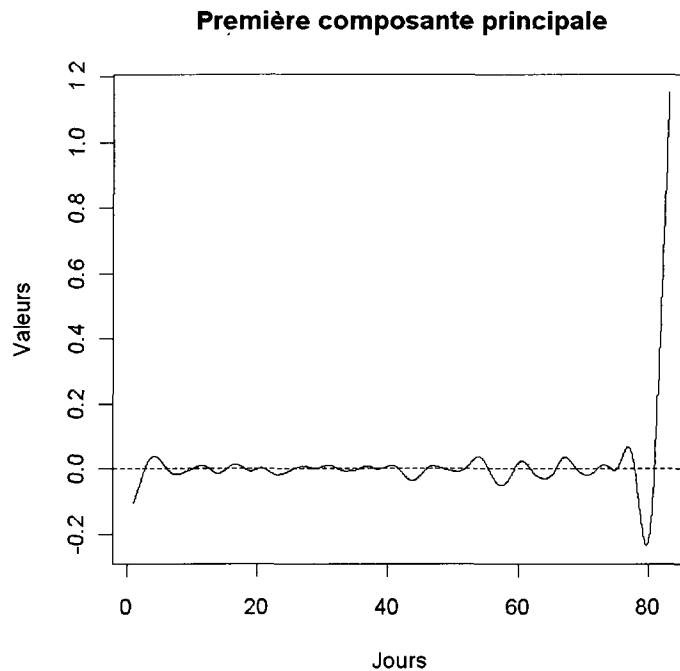


Figure 38 : Valeurs de la première composante principale pour L'ACPF de la deuxième dérivée des courbes de consommation cumulée

4.7 Analyse discriminante fonctionnelle de données de prises alimentaires

Dans cette section, on tente de déterminer s'il est possible de discriminer parmi deux groupes, soit les porcs en bonne santé et les porcs qui ont connus un ou plusieurs épisodes de fièvre durant la période d'observation, en utilisant les données cumulées de consommation alimentaire porcine. Avant de mettre en œuvre l'analyse discriminante fonctionnelle, telle que décrite au chapitre précédent, il faut déterminer lequel des critères, parmi ceux du tableau 16 (section 4.3), on devra utiliser et la nature des courbes qui, en termes statistiques, ont le meilleur potentiel discriminant. Ce travail est effectué dans la sous-section suivante.

4.7.1 Détermination du critère de discrimination et de la nature des courbes à utiliser

Tout d'abord, il faut rappeler qu'une analyse discriminante fonctionnelle de données (appartenant à deux groupes distincts), comme dans cette application, s'apparente à l'analyse en composantes principale fonctionnelle, car c'est une ACPF des centres de gravité ou des moyennes de ces deux groupes en utilisant une métrique particulière. Donc, pour déterminer le critère et la nature des courbes à utiliser, on présentera une série de figures illustrant les courbes de consommation moyenne et de ses dérivées pour certains critères. En procédant de cette façon, on espère arriver à des résultats compréhensibles, significatifs et applicables pour les experts en production porcine.

Tout d'abord, la figure 39 présente les moyennes des courbes de consommation alimentaire cumulée des porcs malades (ligne rouge pointillée) et des porcs en santé (ligne noire pleine) selon les sept critères du tableau 16 (section 4.3).

À partir de ces graphiques, on remarque que, pour chacun des critères, la courbe représentant la moyenne de la consommation cumulée est pratiquement toujours sous celle représentant la moyenne de la consommation cumulée des porcs pour lesquels on n'a jamais observé un épisode de fièvre. Par contre, il est impossible de déterminer visuellement lequel des différents critères maximise l'écart entre les deux courbes moyennes. Pour se faire, on utilise la somme des écarts, en kg, entre les courbes des porcs en santé et celle des porcs malades évaluées en chacun des nœuds. Ces résultats sont illustrés à la figure 40. Il est important de noter que l'on calcule bien l'écart et non l'écart en valeur absolue entre ces courbes pour défavoriser les critères où les moyennes se croisent à répétition.

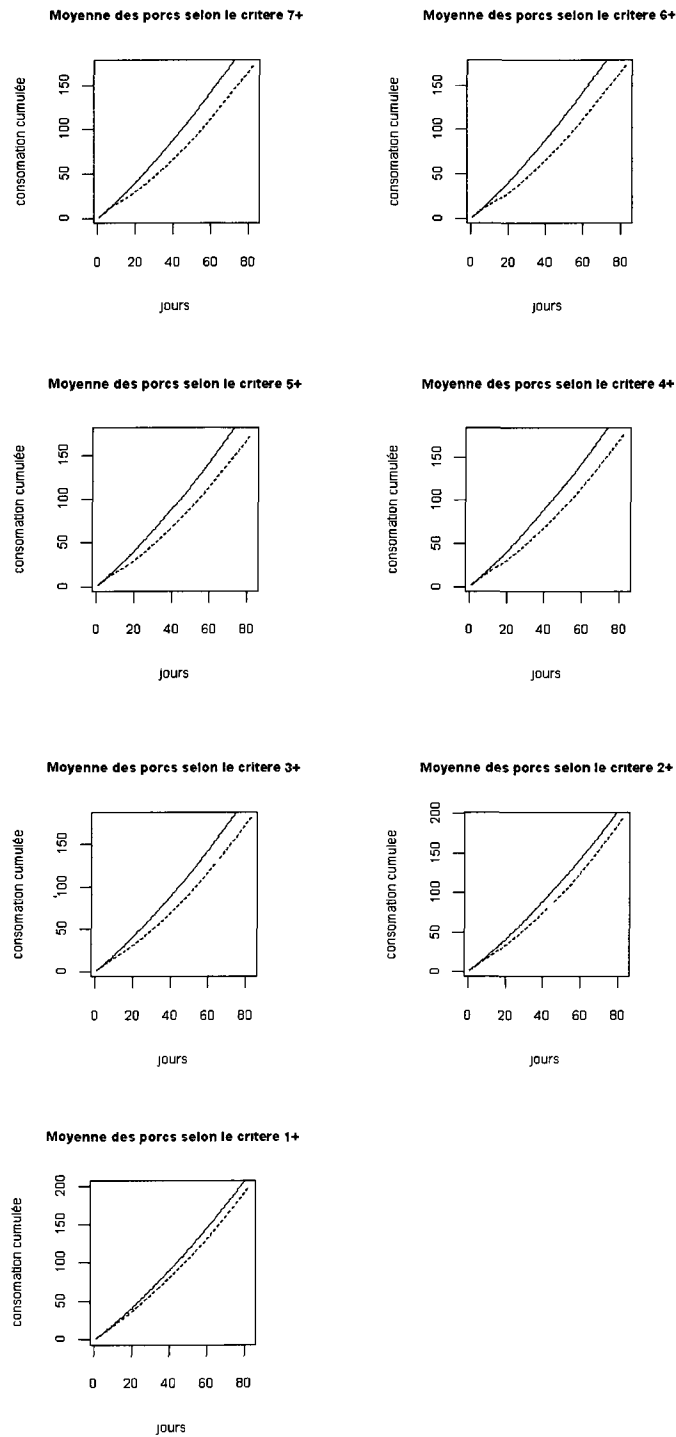


Figure 39 : Moyenne des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les différents critères

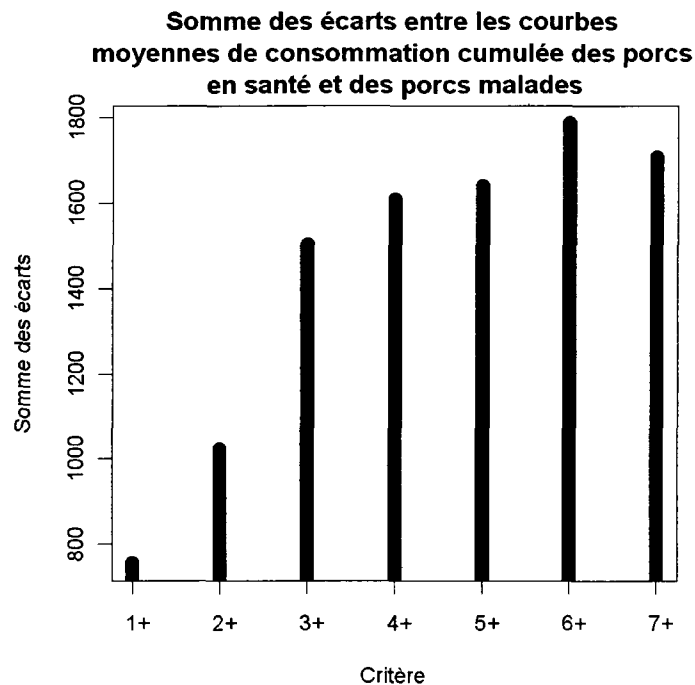


Figure 40 : La somme des écarts entre les courbes moyenne de consommation cumulée des porcs en santé et des porcs malade selon les sept critères

À partir de cette dernière figure, on remarque que le critère qui dit qu'il faut au moins six jours consécutifs de fièvre pour déclarer un porc malade est celui qui semble, du point de vue statistique, avoir le plus grand potentiel discriminant. Donc, c'est ce critère que l'on sélectionne pour effectuer l'analyse discriminante fonctionnelle des courbes de consommation cumulée. De plus, comme la différence, en termes de la somme des écarts, entre le critère 6+ et 3+ est petite, on tentera aussi une analyse discriminante selon ce critère dans le but de pouvoir discriminer le plus tôt possible les porcs malades des porcs en santé.

Comme, on l'a fait pour l'ACPF, il est possible d'analyser la dérivée première des courbes de consommation cumulée. Donc, la figure 41 et 42 illustre la moyenne de ces dérivées selon les sept mêmes critères.

On remarque, en regardant ces graphiques que de la moyenne la dérivée première que cette mesure semble avoir un certain pouvoir discriminant et cela pour l'ensemble des critères, car on remarque que les courbes représentants le groupe des porcs malades, toujours les lignes rouges pointillées, sont presque toujours sous celles du groupe des porcs en santé. Pour savoir lequel des critères on gardera, on calcule encore la somme des écarts en les courbes qui sont présenté à la figure 43.

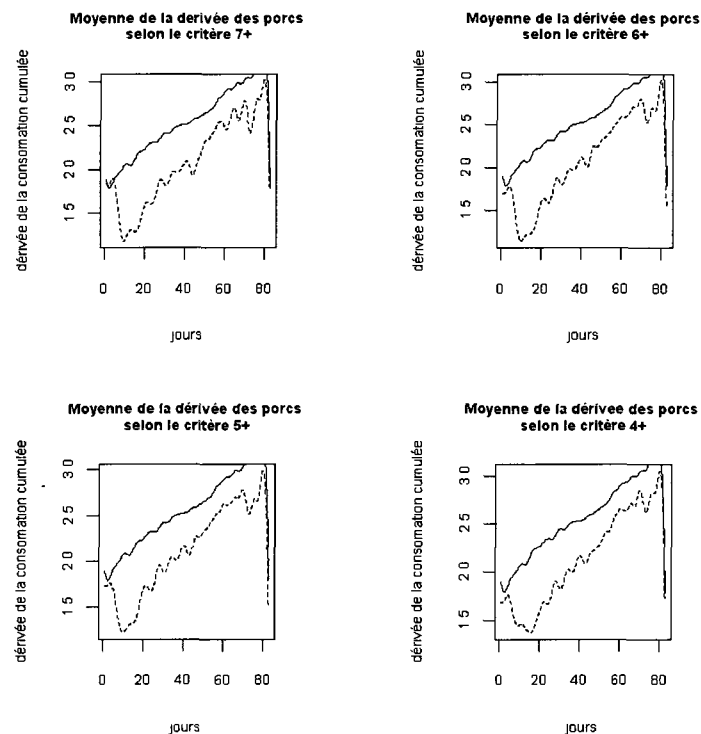


Figure 41 : Moyenne de la première dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les différents 4+ à 7+

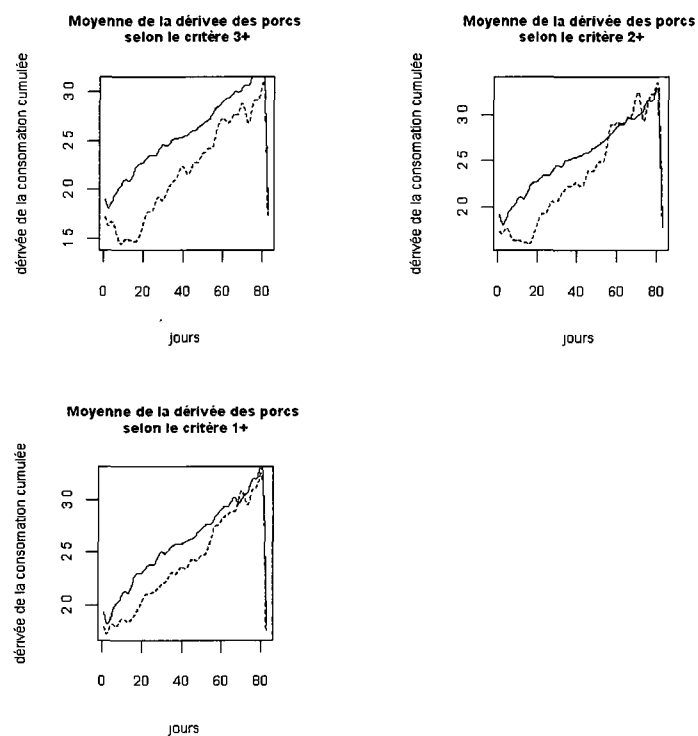


Figure 42 : Moyenne de la première dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 1+ à 3+

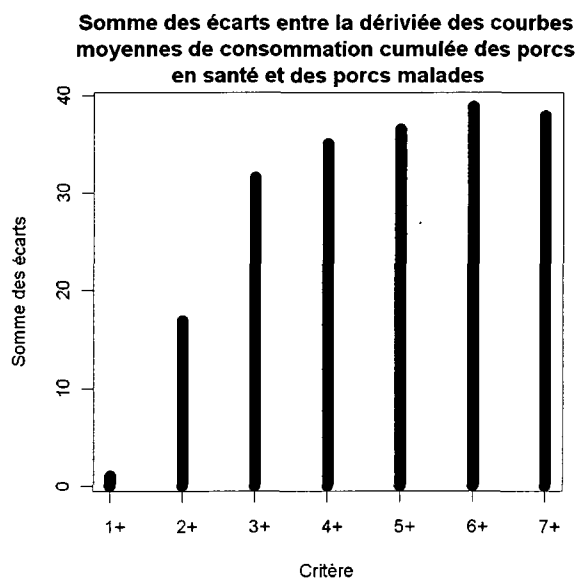


Figure 43 : La somme des écarts entre la première dérivée des courbes moyennes de consommation cumulée des porcs en santé et des porcs malades selon les sept critères

On peut voir, par la figure 43 que le critère qui consiste à classer un porc dans le groupe des porcs malades si l'on a observé plus de six jours consécutifs de fièvre, est celui pour lequel l'écart calculé est plus grand. Donc, c'est ce critère qui est retenu pour l'ADF de la première dérivée des courbes de consommations alimentaire cumulée.

Pour conclure cette sous-section, on porte notre attention sur la deuxième dérivée des courbes de consommation cumulée. On remarque, en regardant la figure 44 et 45, que la dérivée deuxième n'est pas une fonction qui a un fort pouvoir discriminant entre les porcs malade, ligne rouge pointillée, et en bonne santé, ligne noire pleine, car les moyennes sont très similaires et ne cessent de se croiser. Pour ces raisons et aussi parce les courbes de consommation cumulée et leurs dérivée semblent plus prometteur, on décide de ne pas effectuer un analyse discriminante fonctionnelle sur la dérivée deuxième des courbes de consommation alimentaire cumulée.

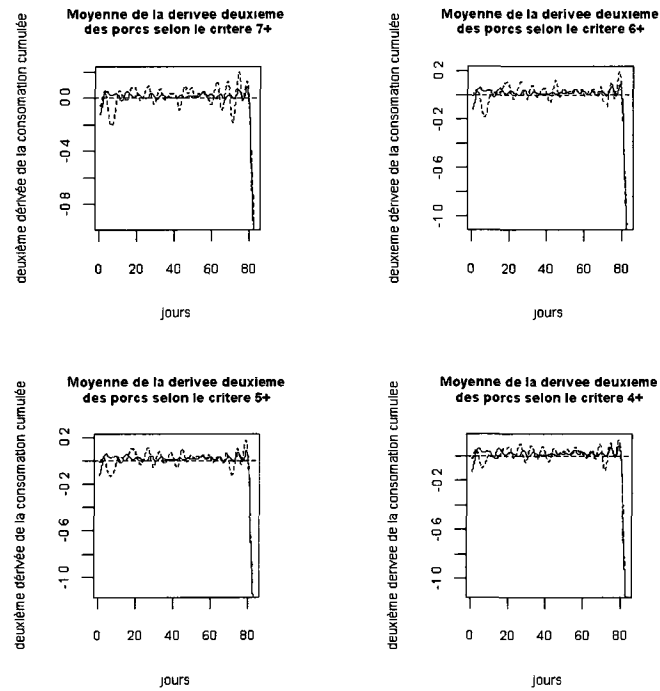


Figure 44 : Moyenne de la deuxième dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 4+ à 7+

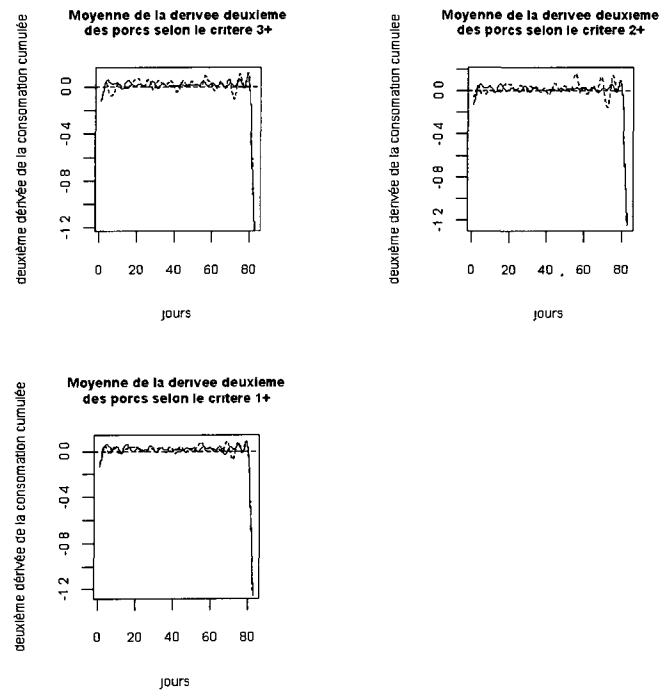


Figure 45 : Moyenne de la deuxième dérivée des courbes de consommation cumulée pour les porcs malades et les porcs en bonne santé selon les critères 1+ à 3+

4.7.2 Application de l'analyse discriminante fonctionnelle

Un fois le critère de discrimination choisi, on peut effectuer l'analyse discriminante. On commence par l'appliquer aux courbes de consommation cumulée selon le critère 6+, soit qu'un porc est classé dans le groupe des porcs malades s'il a eu au moins six jours consécutifs de fièvre. La figure 46 présente les courbes de consommation cumulée des 10 porcs malade (lignes noire pleine) et de la fonction discriminante (ligne rouge pointillée) sous ce critère.

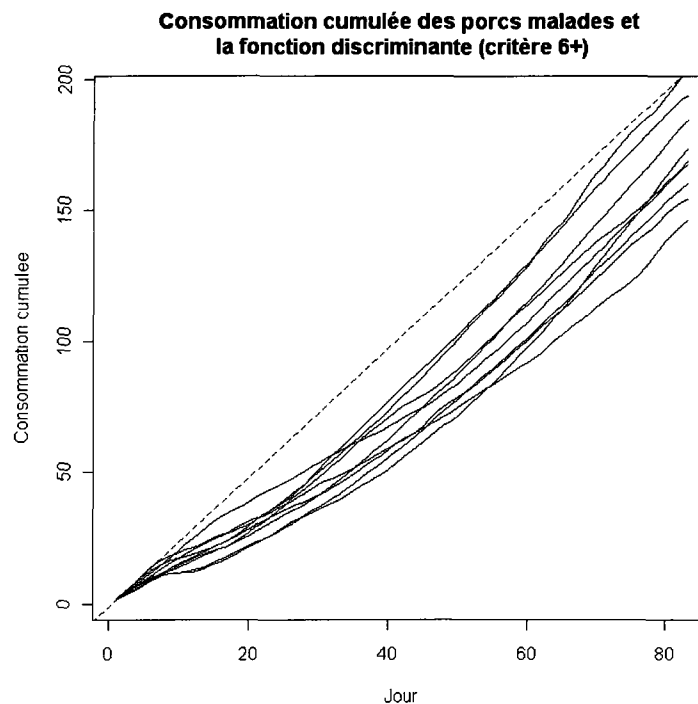


Figure 46: Consommation cumulée des porcs malades et la fonction discriminante sous le critère 6+

On remarque que les courbes de consommation cumulée pour les 10 porcs malades sont presque toujours sous la fonction discriminante ce qui indique un fort potentiel de discrimination entre le groupe des porcs malades et en santé de ce modèle.

On doit aussi quantifier l'efficacité du modèle. Pour ce faire, on regarde le pourcentage d'observations bien classées, c'est-à-dire, que l'on applique la règle de discrimination qui a été calculée dans le modèle à l'ensemble des observations, ce qui donne la classification *a posteriori* des observations que l'on compare avec la classification *a priori*. Pour une observation donnée, si les deux classifications sont identiques, on dira de cette observation qu'elle est bien classée. La figure 47 présente la classification *a posteriori* de l'ensemble des observations

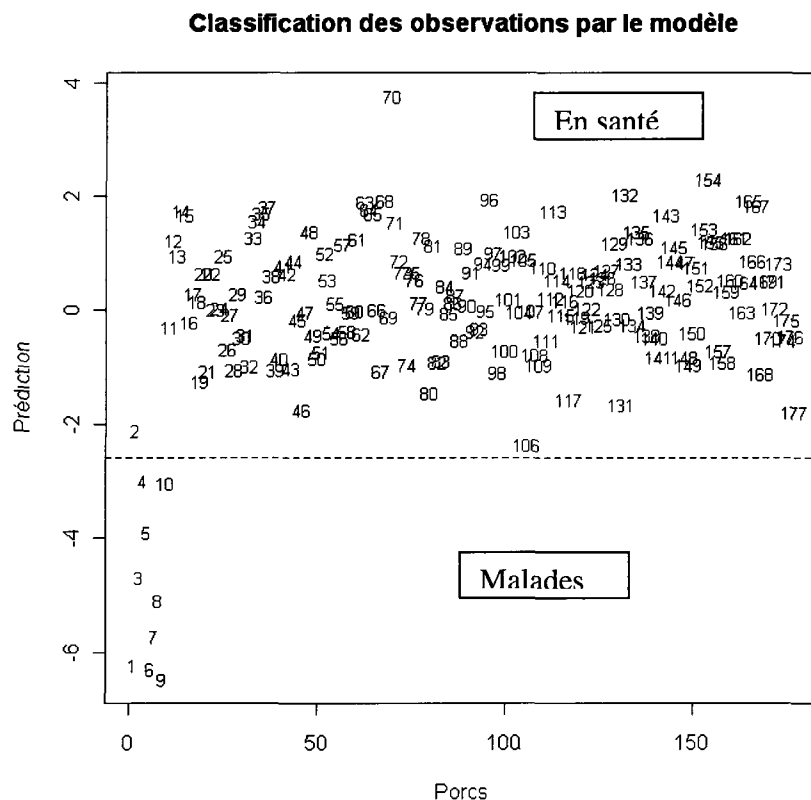


Figure 47 : Classification *a posteriori* des observations selon le critère 6+

À la figure 47, les observations en rouge sont celles qui ont été classées dans le groupe des porcs en santé et celles qui ont été classées dans le groupe des porcs qui ont été malades. Selon la classification *a priori*, les porcs 1 à 10 sont

des porcs malades. Seulement une observation, le porc numéro 2, a été mal classée ce qui laisse 90% de bien classées dans le groupe des porcs malades et 100% des observations bien classées pour le groupe des porcs en santé. Donc, on peut dire que ce modèle, appliqué sur les courbes de consommation cumulée est efficace pour discriminer entre les porcs malades et en santé.

On peut aussi appliquer l'analyse discriminante aux dérivées premières des courbes de consommation cumulée et cela toujours pour le critère 6+. La figure 48 présente la classification a posteriori du modèle d'analyse discriminante appliqué à ces courbes.

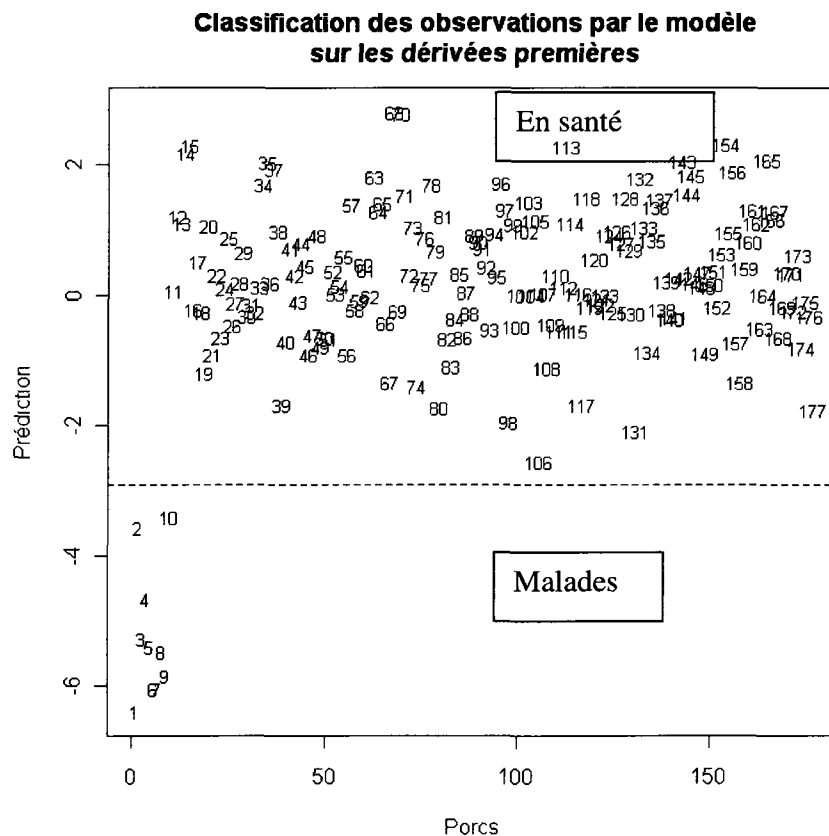


Figure 48 : Classification a posteriori des dérivées premières selon le critère 6+

Dans cette dernière figure, les observations en rouge sont celles classées dans le groupe des porcs en santé et celles en noire dans le groupe des porcs malades. Comme les porcs 1 à 10 sont encore les porcs qui appartiennent *a priori* au groupe des porcs malade, on remarque 100% des observations ont été bien classées. Donc l'analyse discriminante des premières dérivées des courbes de la consommation cumulée sous le critère 6+ est un meilleur modèle que le précédent, en termes de pourcentage observations bien classées, pour départager les porcs malades des porcs en santé.

Maintenant que l'on connaît l'efficacité de l'analyse discriminante pour ce type de données, particulièrement dans le cas de la dérivée première, on croit bon d'effectuer un dernier modèle, mais cette fois, avec un critère différent. On opte pour le critère 3+ qui signifie d'un porc est classé *a priori* dans le groupe des porcs malades s'il a eu plus de trois jours consécutifs de fièvre. L'avantage de ce critère est qu'il permet de déclarer un porc malade plus rapidement, comparativement au critère qui est basé sur six jours consécutifs. Une détection rapide de la maladie suivie d'un isolement du porc en question peuvent diminuer la contagion dans l'ensemble de la production.

On effectue donc une analyse discriminante sur la dérivée première des courbes de consommation cumulée sous le critère 3+ de classification *a priori*. Sous ce critère, 19 porcs sur les 177 ont été classifiés dans le groupe des porcs malades. La figure 49 présente, de façon similaire aux autres analyses, la classification *a posteriori* de l'ensemble des observations.

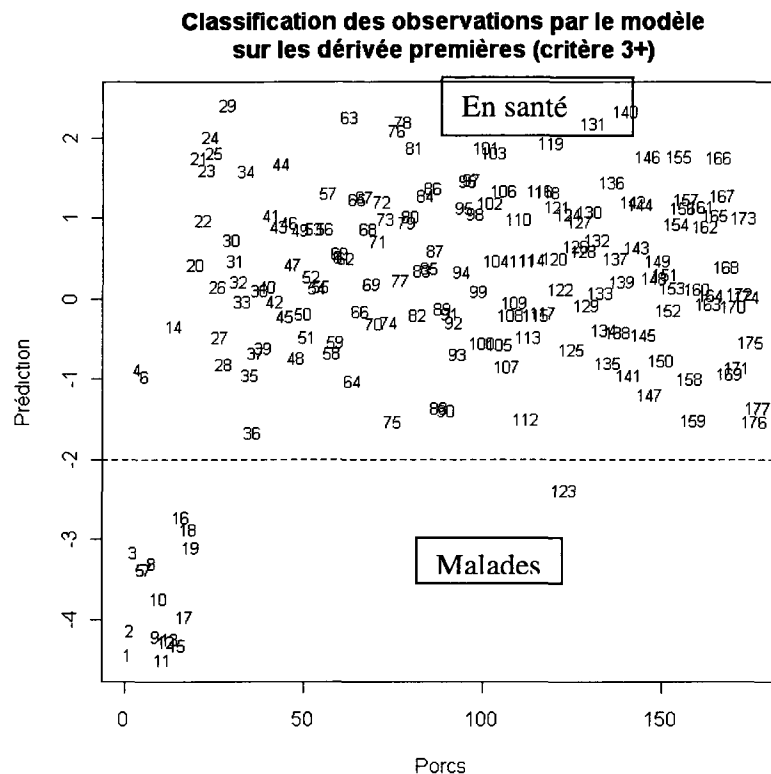


Figure 49 : Classification *a posteriori* des dérivées premières selon le critère 3+

On voit que ce modèle est tout de même efficace, car seulement quatre observations ont été mal classées. Les porcs 4, 6 et 14, *a priori* des porcs malades, ont été classées avec les porcs en santé et le porc 123 se retrouve dans la situation inverse. Malgré cela, près de 85% des porcs malades ont été détectés pour un pourcentage total d'observations bien classées de 98%.

Pour conclure cette sous-section, on peut affirmer que l'analyse discriminante fonctionnelle de données de consommation alimentaire porcine est un modèle convenable pour distinguer les porcs malades des porcs en santé selon les critères retenus. Par contre, il faudrait soumettre cette méthode à des tests sur d'autres échantillons pour évaluer si elle est robuste.

Conclusion

L'analyse de données fonctionnelles est une branche de la statistique qui est en pleine expansion. Cela est imputable à l'augmentation des données disponibles et aussi à la précision des instruments de mesure disponibles. L'évolution des outils informatiques utilisés pour effectuer ce type d'analyse a aussi un effet positif sur l'accessibilité des méthodes.

Ce mémoire avait pour but d'utiliser des méthodes de la statistique conventionnelle telle l'analyse en composantes principales ou l'analyse discriminante et de les appliquer à des données fonctionnelles. Un effort a été fait pour démontrer les bases théoriques entre les deux approches.

De plus, les méthodes présentées ont été appliquées à des données de prise alimentaires quotidiennes cumulées. Cet exemple a permis de voir que l'analyse en composantes principales de telles données permet de représenter les observations dans un espace de dimension nettement plus petit que l'espace d'origine et cela tout en gardant la grande majorité de l'inertie contenue dans les données.

Finalement, on a pu montrer que les courbes de consommation cumulée peuvent servir à détecter si un porc est malade ou en santé en effectuant une analyse discriminante. Les résultats sont plus que convenables, car dans les trois analyses qui ont été effectuées, le pourcentage de porcs, a priori, malades qui ont été bien classés varie entre 85% et 100%.

Bibliographie

- Agriculture and Agri-Food Canada (AAFC), (1993), Recommended code of practice for the care and handling of farm animals-pigs.
- Aguilera A.M., Ocana F.A., Valderrama M.J., (1997), An approximated principal component prediction for continuous time stochastic processes. *Applied Stochastic Models and Data Analysis*, 13, pp. 61–72.
- Aubin J-P., (2000), *Applied functional analysis*, second edition, New York: Wiley-Interscience.
- Baccini A., Besse P., (1999), *Statistique descriptive multidimensionnelle*, Publication du laboratoire de Statistique et Probabilités, UMR CNRS C5583, Université Paul Sabatier.
- Barra V., (2002), Étude de données d'expression par analyse en composantes principales fonctionnelles, *Journées ouvertes biologie informatique mathématique*, p 343-347.
- Benzecri, J. P., (1969). Statistical analysis as a tool to make patterns emerge from data. *Methodologies of Pattern Recognition* (Ed. by S. Watanabe), Academic Press, New York pp. 35-60.
- Besse P., Ramsay J.O., (1986), Principal component analysis of sampled curves. *Psychometrika*, 51 : pp. 285–311.
- Besse P., (1991), Approximation spline de l'analyse en composantes principales d'une variable aléatoire hilbertienne. *Annales de la Faculté des Sciences de Toulouse*, XII(3) : pp. 329–349.
- Bertier, P., Bouroche J-M., (1975), *Analyse des données multidimensionnelles*, Première édition, Presses Universitaires de France, Collection Systèmes Décisions.
- Biau G., Bunea F., Wegkamp M., (2005), Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51, pp. 2163–2172.
- Bouroche, J.M., Saporta, G. , (1980), *L'analyse des données*, PUF, Col. "Que sais-je ?".
- Bracewell, R.N., (1999) *Convolution Theorem. The Fourier Transform and Its Applications*, 3^e édition New York : McGraw-Hill 1999.
- Brezis, H., (2005), *Analyse fonctionnelle : Théorie et application*, Dunod.

Cardot H., Ferraty F., Sarda P, (1999), Functional Linear Model. Statistics and Probability Letter, 45 : pp. 11–22.

Chenguiti K., (2006), Analyse en composantes principales fonctionnelle : cas des prises alimentaires, Thèses canadiennes, Mémoire de maîtrise, Université de Sherbrooke.

Clarkson D.B., Fraley C., Gu C.C., Ramsay J.O, (2005), S+ bFunctional Data Analysis, User's Manual for Windows®, Springer.

Colmez, P., (2009) Éléments d'analyse et d'algèbre (et de théorie des nombres), Éditions École Polytechnique, 470 pages

Crucianu M., Asselin de Beauville J.P., Boné R, (2002), Méthodes factorielles pour l'analyse des données, Germes Science, Lavoisier.

Dabo-Niang S. , Ferraty F, (2008), Functional and Operatorial Statistics. Eds Physica-Verlag.

Dauxois J., Pousse A. (1976), Les analyses factorielles en calcul des probabilités et en statistique : Essai d'étude synthétique. Thèse d'état, Université Paul-Sabatier, Toulouse.

Dauxois J., Ferré L., Yao A.F., (2001), Un modèle semi-paramétrique pour variable aléatoire hilbertienne. C.R. Acad. Sci. Paris, 327(I), pp. 947–952.

Deville J. C., (1974), Méthodes statistiques et numériques de l'analyse harmonique., Annales de l'INSEE, 15.

Ferraty F., Vieu P., (2006), NonParametric Functional Data Analysis Theory and Practice by Springer Series in Statistics, Springer, New York.

Ferraty F., Vieu P., (2002), The functional nonparametric model and application to spectrometric data. Computational Statistics, 17, pp. 515–561.

Ferré L., Yao A.F., (2003), Functional sliced inverse regression analysis. Statistics, 37, pp. 475–488.

Ferré L., Yao A.F., (2005). Smoothed functional inverse regression. Statistica Sinica, 15(3), pp. 665–683.

Ferré L. Villa N., (2005), Discrimination de courbes par régression inverse fonctionnelle. Revue de Statistique Appliquée, LIII(1) : 39–57.

- Fisher R. A., (1936), The use of multiple measurement in taxonomic problems, *Annals of Eugenics*, vol 7, pp.179-188.
- Gauss C. F. ,(1809), *Theoria motus corporum celestium*. Hamburg: Perthes et Besser.
- Guelfand, I. M., Vilenkin, N. Y. (1967), *Les Distributions: Tome 4. Applications de l'analyse harmonique*, Dunod, Paris, 1967, 376 pages
- Hastie T., Tibshirani T., Buja A., (1994), Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89, pp. 1255–1270.
- Hastie T., Buja A., Tibshirani R., (1995), Penalized discriminant analysis. *Annals of Statistics*, 23 : pp. 73–102.
- Hotelling K., (1933), Analysis of a Complex of Statistical Variables Into Principal Components, *Journal of Educational Psychology* pp. 417-441.
- Ivanov V.K., (1962) On linear problems which are not well-posed. *Soviet. Math. Docl.*
- James, G., and Hastie, T, (2001), Functional Linear Discriminant Analysis for Irregularly Sampled Curves, *Journal of the Royal Statistical Society Series B* 63, 533-550.
- James G.M., Hastie T. J., (2001), Functional Linear Discriminant Analysis for Irregularly Sampled Curves, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 63, No. 3, pp. 533-550.
- Kaiser H., (1958), The Varimax Criterion for Analytic Rotation in Factor Analysis, *Psy-chometrika* 23, pp.187-200.
- Kullback, S., (1959), *Information Theory and Statistics*. John Wiley & Sons, New York.
- Leborgne, D. (1991), *Calcul différentiel complexe*, Presses Universitaires de France, 127 pages
- Lee S-Y., Zhang W., Song X-Y. Estimating, (2002), The covariance function with functional data, *British Journal of Mathematical and Statistical Psychology*, Volume 55, Number 2, pp. 247-26.
- Legendre A. M., (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris : Courcier.

Letourneau Montminy, M.-P., C. Boucher, C. Pomar, F. Dubeau, Dussault J.-P., (2005). Impact de la méthode de formulation et du nombre de phases d'alimentation sur le coût d'alimentation et les rejets d'azote et de phosphore chez le porc charcutier. *Journée Recherche Porcine*, 37, pp. 25-32.

Leurgans S.E., Moyeed R.A., Silverman B.W., (1993), Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B*, 55, pp. 725–740.

Pearson K., (1901), On Lines and Planes of Closest Fit to Systems of Points in Space, *Philos. Mag.* 2, pp. 559-572.

Pezzulli S., Silverman B.W., (1993) Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8, pp. 1–16.

Pomar C., Pomar J., Babot D., Dubeau F., (2007), Effet d'une alimentation multiphase quotidienne sur les performances zootechniques, la composition corporelle et les rejets d'azote et de phosphore du porc charcutier, *Journées Recherche Porcine*, 39, pp. 23-30.

Pomar, C., F. Dubeau, M.-P. Letourneau Montminy, M. Mahé, P.-O. Julien, Jondreville C., (2004). Réduction de l'excrétion de phosphore et d'azote chez le porc charcutier par l'ajout d'un objectif environnemental dans l'algorithme traditionnel de formulation. *Journée Recherche Porcine* 36, pp. 251-258.

Preda C., Saporta G., (2002), Régression PLS sur un processus stochastique. *Revue de statistique appliquée*, L(2).

Preda C., Saporta G., (2005), PLS discriminant analysis for functional data. In *ASMDA 2005 proceedings*, pp. 653–661.

Ramsay J.O., Hooker G., (2009), *Functional Data Analysis with R and MATLAB*, Spencer Graves, Springer New York.

Ramsay J.O., Silverman B.W., (2005), *Functional Data Analysis*, Second edition, Springer Series in Statistics, Springer.

Ramsay J.O., Silverman B.W., (2002), *Applied Functional Data Analysis : Methods and Case Studies*, Second edition, Springer Series in Statistics, Springer.

Ramsay J., Dalzell C.J., (1991), Some tools for functional data analysis (with discussion), *Journal of the Royal Statistical Society, B*, 53, pp. 539-572.

Risler J.-J., (1991), *Méthodes mathématiques pour la CAO*, Masson, Paris.

- Rossi F., Conan-Guez B., El Golli A., (2004), Clustering functional data with the som algorithm. ESANN 2004 proceedings, pp. 305–312.
- Rossi F., Conan-Guez B., (2005), Functional multi-layer perceptron: a nonlinear tool for functional data analysis. *Neural Networks*, 18(1), pp. 45–60.
- Rossi F., Delannay N., Conan-Guez B., Verleysen M., (2005), Representation of functional data in neural networks. *Neurocomputing*, 64, pp. 183–210.
- Rotella F., Borne P.,(1995), Théorie et pratique du calcul matriciel, Thchnip.
- Samuelides M., Touzillier, L., (1990), Analyse harmonique Cépaduès-Éditions, 224 pages
- Samuelides M., Touzillier L., (1989). Analyse Fonctionnelle, Cépaduès-Éditions, 289 pages.
- Saporta, G., (1990), Probabilités, Analyse des données et Statistiques, éditions Tech-nip. 4.
- Saporta G., (1981), Méthodes exploratoires d'analyses des données temporelles. Cahiers du BURO, pp.37-39.
- Schwartz L., (1979), Analyse Hilbertienne, Herman, Paris.
- Silverman B.W., (1996), Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, 24. pp. 1–24.
- Spearman, C., (1904). "General intelligence" objectively determined and measured. *"American Journal of Psychology"*, 15, pp. 201–293.
- Tartar L., (2007), An Introduction to Sobolev Spaces and Interpolation, Springer.
- Tenenhaus M.,(1996), Méthodes statistiques en gestion , Dunod.
- Tihonov A.N., (1963), Regularization of incorrectly posed problems. *Soviet Math. Docl.*, 4 : pp.1624–1627.
- Tihonov A.N., (1963), Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Docl.*, 4 : pp.1036–1038.
- Thurstone, L. L., , Thurstone, T. G., (1941), *Factorial studies of intelligence*. Chicago: University of Chicago Press.
- Wagschal C, (1995), *Topologie et analyse fonctionnelle*, Édition Hermann, Collection Méthodes.